Report about a workshop on sensitive data:

Repositories for sharing individual participant data from clinical trials and existing tools/services for storing clinical trial data

The workshop on sensitive data was performed within the CORBEL project and was held on 14 June 2018 in Paris, France. The report was prepared by the CORBEL WT3.3 project group and covers the following areas:

1. **Background** (from the CORBEL project)
2. **Assessment of existing repositories for sharing individual participant data (IPD) from clinical trials**
   (Study protocol (March 2018), Status report on assessment of data repositories (June 2018))
3. **Survey of existing tools/services for storing clinical trial data**

Grant information

This project has received funding from the European Union's Horizon 2020 research and innovation programme (CORBEL, under grant agreement n° 654248).

1. Background

Within the H2020-funded project CORBEL (Coordinated Research Infrastructures Building Ensuring Life-science Services; see : http://www.corbel-project.eu/home.html), the objective of CORBEL working task 3.3 is to develop procedures to provide the scientific community with access, upon request, to the patient-level data from previous clinical trials for re-analyses, secondary analyses and meta-analyses and to test them in a pilot/demonstrator. As a first step a multi-stakeholder patient-level data taskforce has been established and recommendations for providing access to IPD from clinical trials have been developed in a consensus process and published openly (BMJ Open, 2017). Further work resulted in a detailed, structured and comprehensive list of processes/subprocesses involved and tools/services needed for data sharing (F1000 Research, 2018). Based on high-level requirements and evaluation of existing tools for a data sharing platform, the next step of work foresees the development of a pilot/demonstrator for data sharing. In an internal CORBEL meeting, performed on 26 February 2018 in Paris, it was concluded that more preparatory work has to be performed in order to be able to take the decision for selecting or adapting an existing repository or developing a new pilot/demonstrator. It was suggested to evaluate existing repositories, already storing IPD from clinical trials, and to survey existing tools and services for handling of sensitive data. In a second meeting, held in Paris on 14 June 2018, this was discussed further and the envisaged sub-projects were defined more precisely:

a) Assessment of existing repositories for sharing individual participant data (IPD) from clinical trials

b) Survey of tools and services for clinical trial data

This document specifies the methodology used in these two subprojects and the status of work performed so far (June 2018).

2. Assessment of existing repositories for sharing IPD from clinical trials

2.1 Study protocol (March 2018)

Rita Banzi - Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy.

Steve Canham - Canham Information Systems, Surrey, UK.

Christian Ohmann - European Clinical Research Infrastructure Network (ECRIN), Düsseldorf, Germany.

Wolfgang Kuchinke - Coordination Centre for Clinical Trials, Heinrich Heine University, Düsseldorf, Germany.

Karmela Krleza-Jeric – IMProving Accesss to Clincal Trial data (IMPACT) Observatory, Mediterranean Institute for Life Sciences (MedILS), Split, Croatia.

Serena Battaglia - European Clinical Research Infrastructure Network (ECRIN), Paris, France.

Jacques Demotes-Mainard - European Clinical Research Infrastructure Network (ECRIN), Paris, France.

Outline

Introduction

Research Data repositories (in further text repositories) may be useful tools to effectively and safely share clinical study data. They could potentially increase the quality of archiving, including data curation and standardization, the discoverability of data through the application of metadata schemes, and facilitate the processes of request and transfer of data from generators to users, as well as tracking data utilization [1]. A further issue regards the long-term preservation of data that could be ensured by data repositories supported by appropriate organizational and business models. The number of repositories is growing fast. As of March 2018, Re3data, a global registry of research data repositories, lists more than 2000 data repositories, hosting data and documents from different scientific disciplines [2].

In the context of a consensus exercise on principles and recommendation on how to share and reuse individual-participant data (IPD) from clinical trials, the central role of data repositories to improve data sharing in clinical research has been recognized [3, 4].

Several funders require that data management plan is submitted along with the grant application and some have established dedicated data repositories [5, 6]. Increasingly, medical editors require that data underlying the results presented in journal articles are made available to other researchers and suggest repositories where data can be stored and identified, for instance BMJ, BMC and PLoS Medicine [6-8] The Committee of medical journal editors (ICMJE) recently proposed a common approach to the inclusion of a data sharing plan in the study publication [9].

Some universities and research institutions established data sharing policies and related data archives mainly to improve dissemination and outreach activities. All these initiatives apply to any type of research data and, thus, clinical study data may be potentially stored and accessed through these systems. Several data platforms for sharing clinical study data, Clinicalstudydatarequest.com, Yoda, CEO Cancer/Project Datasphere, have been operating for some years, and others such as Vivli are under development and will be launched soon.

These platforms are not repositories, but they may play an important role in facilitating the interaction between data generators (mostly industry sponsors) and data users.

Clinical researchers and sponsors, especially those from academia, may not be familiar with data management practices and data storage and sharing options. Information about the different repositories that hold clinical research objects (IPD datasets and related documents) may contribute to inform the development of data sharing plan or the choice of where to store data to make them sharable.

Objective

To describe and classify the features of clinical study data repositories that host IPD from clinical studies and inform the following CORBEL next steps:

- Provision of comprehensive information to sponsors and researchers about the data repositories available, including the options and mechanisms for managing data within those repositories, associated costs, and available data access mechanisms.
- Assessment of the need for a demonstrator repository, to store and manage access to individual-participant (IPD) data.
- Specification of requirements and quality indicators for repositories managing IPD from clinical research.

Methods

Repositories

The analysis will be focused on research data repositories developed by public or private institutions that host clinical study data. The focus is on repositories where IPD from clinical study (and associated documents) can be stored and made available to other users. We will exclude clinical data registries/databanks, patient registries, trial registries, repositories developed by pharma companies, data sharing platforms, or broader initiatives as they are not technically data archives. We will focus mainly on repositories covering a broad spectrum of clinical data, but we will be interested in gather information on disease-specific repositories such as those promoted by the NIH-funded institutions.

Through the activities of previous working group within CORBEL and the collaboration with the IMPACT Observatory [10], we have identified over 30 repositories. This list was created by searching the web, reading relevant literature, using Re3data, and from personal contacts and knowledge, and thus may not be fully representative of the totality of data repositories.

The repositories that are the object of this evaluation are listed below:

1. B2Share
2. BioGrid Australia Limited
3. BIRN: Biomedical Informatics Research Network
4. CancerData.Org
5. Clinicalstudydatarequest.com
6. Critical Path Institute

7. CTTI: Clinical Trials Transformation Initiative
8. DataOne
9. Datasphere
10. DRUM
11. Dryad
12. EASY
13. EBCTCG (CTSU)
14. Edinburgh DataShare
15. Fairsharing (formerly biosharing)
16. FDA Janus
17. FigShare
18. FreeBird
19. Global Health Data Exchange
20. Health and Medical Care Archive
21. Henry A. Murray Research Archive at Harvard University
22. ICPSR Inter-university Consortium for Political and Social Research
23. Immune Tolerance Network Trialshare
24. INDEPTH Data Repository
25. IST: International Stroke Trial Database (Edinburgh)
26. Melanoma MMMP- malignant melanoma map project
27. National Addiction & HIV Data Archive Program
28. National Archive of Computerized Data on Aging
29. National Archive of Criminal Justice Data
30. National Data Archive on Child Abuse and Neglect
31. NIDDK Central Repository / National institute of diabetes and digestive and kidney diseases
32. NDAR: National Database for Autism Research (US)
33. Neuroscience Information Framework
34. NIH BioLINCC/Biologic Specimen and Data Repository Information Coordinating Center
35. NIH NIDA: National Institute of Drug Abuse ()
36. NIMH NDCT
37. OSF
38. Pfizer clinical trial data
39. ProAct
40. Roche trials
41. SND: Swedish National Data Service.
42. SOAR Data: Duke Clinical Research Institute
43. SAMHDA: Substance Abuse and Mental Health Data Archive
44. SLMSR: Sylvia Lawry Centre
45. TBI-IMPACT
46. The Knowledge Network for Biocomplexity
47. UFIDR: University of Florida Health Integrated Data Repository
48. UK Data Archive
49. UK Data Service
50. UMIN
51. Vivli
52. YODA: The Yale University Open Data Access Project
53. Zenodo

This list will be revised and expanded in order to characterise all the generic repositories (i.e. those hosting IPD data about different clinical areas) we can identify, and a sample of repositories focusing on specific disease or medical areas.

At a later stage of the project, the current assessment of repositories might be extended to repositories covering genomic or epidemiological data or even health data.

List of items to be assessed

For each repository included in the analysis, we will assess the items reported in Table 1. These features can be grouped in the following six sections:

a) General – main parameters, including scope, funding and establishment date, and description of the repository

b) Data upload and storage – in particular the practical steps required for depositing IPD and related documents, especially data under controlled access. Any potential costs are also included.

c) The data available and how it can be accessed – both quantitative and qualitative descriptions of the types of access provided and of the data, including an indication of the approximate number of clinical studies stored in the repository, as well as the type of material available.

d) Discoverability – the details of metadata schemas applied and how that metadata is made available.

e) Any other relevant points/ comments

f) Provenance and date of the data collected within this project.

Table 1: list of items for data collection

Code	**Name**	**Information sought**
A. General Parameters		
A1	Name	Name of the repository
A2	URL (Website)	URL of the repository's home page, or the most closely related web page.
A3	Location	Physical location(s) of the repository, including host organisation(s)
A4	Self-description, (max. ½ page)	As copied from the web site, edited to highlight main features / claims
A5	Scientific scope	1 = General scientific, 2 = Medical and social sciences, 3 = Clinical and epidemiological studies, 4 = Clinical trials only, 5 = Specified study or group of studies, 6 = Other

A5S	Scientific scope details	Mandatory if 'Other' selected above, optional otherwise. Further details of scientific scope.
A6	Source scope	In terms of the location of the source material, in practice: 1 = Global, no restriction, 2 = Continental, 3 = National, 4 = Regional (within a country), 5 = Institution or Organisation, 6 = Other
A6S	Source scope details	Mandatory if 'Global' not selected above, optional otherwise. Further details of geographic scope including names of nations, regions, organisations etc.
A7	R3data	Indicate Y/N if listed in R3data registry
A7S	R3data URL	The URL of the relevant entry within the R3data system
A8	Start year	The year the repository started (in its current form or something like it)
A9	Main funding	1 = public / academia, 2 = pharma industry, 3 = academia / public and pharma consortium, 4 = Other
A9S	Main funding details	Mandatory if 'Other' selected above, optional otherwise. Further details of main funders.
A10	Repository sustainability	What is the funding position of the repository in the longer term, so far as that can be known?
A11	Business continuity	Is there any commitment to preservation of the data if the repository has to close? (Y/N)
A11S	Business continuity details	If yes to the question above please summarise.
B.	**Data Upload and Storage**	
B1	Rules and guidelines on upload present?	Does the repository have and apply rules and/or guidelines for uploading data? (Y/N)
B1S	Standards and guidelines location	If yes to the question above, the URL or other location indicator for the relevant documents.
B2	Who can upload data?	Any restrictions on who can upload, including details on how the restrictions are applied (e.g. the criteria used for approval)
B3	Format, metadata and documentation requirements for upload?	Are there particular ways in which data must be formatted, and / or metadata and documentation provided? (Y/N)
B3S	Format and doc. Details	If yes to the question above, a summary of requirements.
B4	De-identification practices before upload?	Are there particular requirements or guidelines relating to the de-identification of uploaded data? (Y/N)
B4S	De-identification details	If yes to the question above, a summary of requirements.
B5	Control of quality of data?	Are there control / review mechanisms in place to check data quality when it is submitted to the repository? (Y/N)

B5S	Control of quality details related to upload	If yes to the question above, a summary of the mechanisms in place.
B6	Acceptable size of files per trial per trial	Is there a limit to the size of files and datasets that can be uploaded. (Y/N)
B6S	Acceptable file size details	If yes to the question above, a summary of applicable limits.
B7	Costs of upload?	Is there any cost associated with uploading data, to the data generator or depositor (Y/N)
B7S	Costs of upload, details	If yes to the question above, a summary of possible costs.
B8	Formal contract regarding upload and storage?	Is there a formal agreement to be signed by the data generator and repository specifying the roles and responsibilities of each? (Y/N)
B8S	Formal contract, details	If yes to the question above, a summary of requirements.
B9	Costs of storage	Is there any cost associated with maintaining data in the repository, to the data generator (Y/N)
B9S	Costs of storage, details	If yes to the question above, a summary of possible costs.
B10	Length of storage	Are there any limits on the storage period of uploaded data, or is it viewed as indefinite?
C. Data Available and Access		
C1	Has public clinical study data (IPD) (without self-attestation)?	Whether has clinical study data that can be accessed and downloaded by the public, *without any user self-attestation;* 1 = Yes, 2 = No, and would not, 3 = No but might in the future, 4 = Unknown
C2S	Public access details	Details of public access scheme (e.g. type of data available under that access, any time embargoes operating etc.)
C2	Has public clinical study data (following self-attestation)	Whether has clinical study data that can be accessed and downloaded by the public, *following web-based user self-attestation;* 1 = Yes, 2 = No, and would not, 3 = No but might in the future, 4 = Unknown
C2S	Self-attested access details	Details of web-based self-attestation scheme in operation or planned (e.g. type of data available, attestation details required, confirmation mechanisms).
C3	Has managed access clinical study data?	Whether has clinical study data with controlled access to downloadable files, e.g. through group membership or case by case review; 1 = Yes, 2 = No, and would not, 3 = No but might in the future, 4 = Unknown
C3S	Managed access details	Details of managed access scheme in operation or planned, (e.g. type of data available, mechanisms of proposal review, used of advisory groups)
C4	Has managed access to analysis environment?	Whether has clinical study data that can only be accessed in situ for analysis purposes

		1 = Yes, 2 = No, and would not, 3 = No but might in the future, 4 = Unknown
CS	Analysis environment details	Details of analysis environment in action or planned
C5	Total number of clinical studies involved (at the date of data extraction)	If the repository has clinical study data (B1, B2, B3 or B4), approximately how many studies have published such data in a given repository?
C6	Has other clinical study data objects?	Whether the repository includes other clinical study documents and data, e.g. protocols, CSRs, analysis plans etc.; 1 = Yes, 2 = No, and would not, 3 = No but might in the future, 4 = Unknown
C6S	Types of other data objects, description	A textual description of the various types of data and documents stored within the repository, if there is any specialisation. This includes both types of datasets and types of other data objects. Otherwise indicate that all types of data and documents would be allowed.
D. Discoverability		
D1	Application of an identifier	Are identifiers (e.g. a DOI) assigned to datasets and other material stored within the repository? 1 = Yes, assigned before upload, 2 = Yes, assigned after upload, 3 = Mixed, some material has identifiers, some not, 3 = No, 4 = Unknown
D1S	Details of identifier scheme(s)	If question above answered 1, 2 or 3, indicate the type(s) of identifier used and the responsibilities for assigning them.
D2	Application of a metadata schema to describe its contents	Does the repository use a consistent metadata schema (or schemas) to describe its contents, either internally, on web pages, or both? (Y/N)
D2S	Metadata schema details	If yes to question D2, summarise, or if a known schema please name it, or provide a URL to a description of the schema.
D2T	Metadata schema responsibilities	If yes to question D2, summarise how the metadata schema is applied, when and who by.
D3	Metadata availability on the web	Does the repository provide a web-based catalogue of its contents, searchable by humans (a so called metadata repository)? (Y/N)
D3S	Metadata searchability	If Yes above, indicate the type of filters / criteria that can be used (e.g. keyword, study title, study Id, doi, disease area etc.)
D4	Metadata availability through an API	Does the repository provide an API for machine-based interrogation of its metadata / catalogue? (Y/N)
D4S	Metadata API, location of description	If yes above, give a URL or other location indicator for a description of the API, if available
D5	Access details available for secondary users	Are guidelines / policies about (controlled) secondary access available to users, for example how to apply and to whom? (Y/N)

D5S	Access details, location	If yes above, give the URL for the access details
D6	Costs to user?	Does gaining access to data or other data objects involve a cost to the secondary user, at least in some cases?
D6S	Costs to user details	If yes above, provide details.
D7	Format of accessed data	Is data downloaded or accessed in situ only available in particular formats? 1 = No, data downloaded / accessed is unchanged from original uploaded format, 2 = Yes, data available in particular formats only, 3 = Other
D7S	Formats used	If 2 or 3 above, provide details. If appropriate include software specific formats (e.g. ‘SAS transport files only’, ‘SDTM structured files’) as well as format type (e.g. ‘csv files’).
E. Other Features		
E1	Other features	Any other important information about the repository, with particular focus on CORBEL information needs, not recorded elsewhere.
F. Data Provenance		
F1	Source of info	Research team member (s) that collected /extracted the information, as initials
F2	Date of Info	The last date on which the information was known to be accurate.
F3	Repository contacts	Email address(es) of individuals who are willing to serve as contact points in the future (for this exercise, not publicly)

Data collection

Data will be collected from repository public websites, publications, and other publicly available materials by two independent authors. At this stage, we will not contact repositories to seek for additional information or clarification. Discrepancies will be discussed and resolved.

We will use a data collection form prepared in Access 2016.

Data analysis

The collected data will be described using tables and narrative formats. We will try to build an indicator dedicated to the assessment of the maturity of the repository for hosting clinical study data. The items included in the indicator will be agreed among the authors, and will include for instance funding sustainability, guidelines for data upload, storage, and data de-identification, data quality controls, application of identifiers, and availability of metadata.

Output

- An internal CORBEL document to inform the development of the next steps of the project (deadline beginning June 2018, presentation of interim report at CORBEL meeting, 14 June 2018).
- Explore the possibility to include a dedicated area in the ECRIN website, i.e. database of repositories.
- Preparation of a possible publication summarizing the results during summer-fall 2018.

References

1. Austin CC, Brown S, Fong N. et al Research Data Repositories: Review of Current Features, Gap Analysis, and Recommendations for Minimum Requirements IASSIST Quarterly IASSIST Quarterly Preprint. 2015 International Association for Social Science, Information Services, and Technology

2. R3data https://blog.datacite.org/re3data-science-europe/.

3. Ohmann C, Banzi R, Cahnam S, et al Sharing and reuse of individual participant data from clinical trials: principles and recommendations BMJ Open 2017;7:e018647. doi: 10.1136/bmjopen-2017-018647

4. Krleža-Jerić K., Sharing of clinical trial data and research integrity. Periodicum Biologorum. 2014 116(4):337-339.

5. NIH Sharing Policies and Related Guidance on NIH-Funded Research Resources https://grants.nih.gov/policy/sharing.htm

6. Wellcome trust Policy on data, software and materials management and sharing https://wellcome.ac.uk/funding/guidance/policy-data-software-materials-management-and-sharing

6. BMJ data sharing policy https://authors.bmj.com/policies/data-sharing/

7. BioMed Central data sharing policy https://www.biomedcentral.com/about/policies/open-data

8. PLoS Medicine Data availability https://journals.plos.org/plosmedicine/s/data-availability

9. Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, et al. Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors. PLoS Med 2017 Jun 5;14(6):e1002315.

10. Krleza-Jeric K, Gabelica M, Banzi R, et al. IMPACT Observatory: tracking the evolution of clinical trial data sharing and research integrity. Biochem Med (Zagreb) 2016;26(3):308-07. doi: 10.11613/BM.2016.035.

2.2. Status report on assessment of data repositories (June 2018)

Presented by Rita Banzi on 14 June 2018 - Paris, France ECRIN headquarter

Previous steps

1. March 2016, Paris (first meeting of the CORBEL consensus exercise): establishment of a subgroup responsible for an environmental scan of clinical study data repositories

   Task: to do a review of characteristics of existing data sharing solutions and to identify gaps, if there are any, between those solutions and the needs of the research community.

   Deliverable: document summarizing the existing data sharing solutions and a gap analysis the identified gaps in current functionality of missing features or functions (end of June 2016)

2. March 2017, Paris (third meeting of the CORBEL consensus exercise):
   a. presentation of draft results on 14 generic repositories that host clinical trial raw (IPD) data and are in the public domain.
   b. identification of a subset of disease-specific repositories to be analysed.
3. December 2017, Paris (CORBEL WT 3.3 meeting): discussion of the draft manuscript and decision to revise the main objectives and conduction of the analysis.
4. January – February 2018 (email/teleconferences): definition of the objectives as follows:
   a. Provision of comprehensive information to sponsors and researchers about the data repositories available, including the options and mechanisms for managing data within those repositories, associated costs, and available data access mechanisms.
   b. Assessment of the need for a demonstrator repository, to store and manage access to individual-participant (IPD) data.
   c. Specification of requirements and quality indicators for repositories managing IPD from clinical research
5. March 2018, Paris (CORBEL WT 3.3 meeting): discussion and agreement on a study protocol the
6. March – June 2018:
   a. Refinement of the list of items to be collected,
   b. Elaboration and piloting of the data collection and overall process.
   c. Data extraction concluded
   d. Data reconciliation ongoing

Next steps

- Finalize data reconciliation
- Summarize and analyse data
- Repositories described qualitatively
- use of quality metrics (e.g. scores) rather than indicators
- Internal report (September 2018)
- Publication in a scientific journal

3. Survey of existing tools/services for storing clinical trial data

Wolfgang Kuchinke and Birol Tilki

Coordination Centre for Clinical Trials

Heinrich-Heine University

Düsseldorf, Germany

Introduction

Part of the original description of work (WT 3.3.3) was the development of a pilot/demonstrator, covering database platform, embedded tools, procedures, extraction, modules, interfaces, data sources, user interfaces, platform tested and evaluated and use of ELIXIR-led virtual access solutions. Such a pilot/demonstrator should be based on a high-level requirements engineering process (based upon a consensus document developed by a multi-stakeholder taskforce) and based on decisions on data standards (e.g. CDISC, FDA) and technical specifications for software development using solutions developed in other projects (e.g. EHR4CR, Transform, EUDAT, ELIXIR, TranSMART) as well as the evaluation of tools for the data storing platform.

This subproject was designed to provide an overview on existing tools/services involved in the management of sensitive data, with the goals of exploring which tools/services were already used for the sharing of individual participant data from clinical trials and which tools/services (singly or in combination) could be used for developing repositories. The methodological approach and the results of this subproject are presented in this document.

Purpose of this work

The purpose of this subproject was to conduct a broad analysis of solutions and tools available for storage and sharing of individual participant data from clinical trials. Because clinical trials data is sensitive data, we had to extend our research to solutions that can deal with sensitive data in a legally compliant manner. Nonetheless, we focused on solutions that had been developed in EU projects and were Open Source. Based on the requirements of the stakeholder group we assumed that it is necessary to cover data management, metadata management, access management, encryption, anonymisation as functional components of a solution.

Methods

Based on requirements, which had been identified implicitly in the consensus document produced by the CORBEL stakeholder group [1], and then explicitly by later work on data sharing processes and actors [2], we searched for already existing solutions that may be suitable to store and provide access to individual participant level data from clinical trials. Because clinical trial data is sensitive data, the requirement to deal with restricted access, data privacy protection and secure storage conditions would seem to exclude all repositories that are adapted to deal only with open data and open access. But many general data repositories provide already some kinds of identity management and authorisation for restricted access. We therefore searched also for additional tools, like anonymisation software that might be used to adapt a data repository to deal with clinical trials data in a compliant way.

We collected information by searching the internet, publications, group reports, as well as in the output of EU projects, like ELIXIR, BioMedBridges, EUDAT and Research Data Alliance (RDA). We used results of the "EUDAT Group for Sensitive Data"[1] and especially the information that was gathered in three workshops where experiences of 20 research communities with their solutions for the collection, processing and analysis of sensible data was discussed [9, 10].

Also included were solutions that were recommended by experts during meetings of the CORBEL, EUDAT and RDA groups. We searched mainly for software solutions that are available as open source, but also included a small number of commercial solutions where they seemed innovative. A template for the collection of the collected information was created with following criteria:

- name of solution / tool
- contact address / person, web page of the tool, country
- possibility for cross-country use
- short description of the tool
- type of activity the tool supports (project, service, collaboration, platform, etc.)
- modules / architecture / components
- kind of data stored
- research use cases / projects / studies the tool is already used (including collaborations)
- comments

Results

[1] https://www.eudat.eu/a-eudat-working-group-on-sensitive-data-management

Stakeholder requirements

The requirements[2] of the CORBEL stakeholder group were extracted from “Sharing and reuse of individual participant data from clinical trials: principles and recommendations”, a paper that summarises the outcome of three stakeholder workshops [1]. The requirements are listed and attached as Annex 1.

Results of the survey

We collected altogether 54 solutions / tools connected with the storage of clinical trials data. The complete list with the results is attached as Annex 2 to the report. During analysis of the results, we saw it was convenient to structure the data and simplify the presentation of the results by grouping them as follows:

- Generic Repository Systems
- Specific Repository Systems
- Group of complementary tools and solutions
    - Generic tools and solutions
    - Specific tools and solutions

Generic Repository Systems

This group consists of large data repository systems, among others, DSpace, Samvera, Hydra-In-a-Box, Fedora, Clinical Data Repository (CDR), MongoDB and ownCloud. These solutions are used to create complex repositories complete with access control, metadata registration, back-up etc. and are already used by many institutions and projects. Especially DSpace and Fedora have been applied widely for institutional data repositories. Further solutions are OwnCloud, claiming to be the leading Open Source Cloud Collaboration Platform. An example of an institution using a repository system like DSpace for clinical data is the Clinical Data Repository of the University of Minnesota (UMN), which stores data from electronic health records (EHRs) of more than 2 million patients from 8 hospitals and more than 40 clinics. For each patient, patient data is available regarding demographics (age, gender, language, etc.), medical history, allergies, immunizations,

[2] See: Schuengel, Manuela. (2016). CORBEL Communication Strategy based on stakeholder analysis. Zenodo. DOI: 10.5281/zenodo.154033

outpatient vitals, diagnoses, medications, results of lab tests, visit locations, providers, etc. Data in this repository can be used for biomedical research, including recruitment planning, retrospective cohort studies, and observational studies. This example shows that it is not necessary to import patient data into a data warehouse to use it for research or recruitment purposes. Recently, UMN has made investments for new infrastructure to support collaboration and data storage totalling $12M to build this research infrastructure. A value of over 10M Euro for the development of a research infrastructure for clinical data was confirmed during stakeholder discussions. ECRIN should be prepared to require a substantial amount of money in case a sustainable repository should be developed and maintained.

MongoDB is a NoSQL document database with scalability and flexibility. It has more than 4,900 customers and stores data in JSON-like documents. MongoDB Stitch is a back-end service that provides an HTTP API to integrate MongoDB with other services. The MongoDB Connector for Business Intelligence (MDB BI) allows users to create queries with SQL, visualize and report using existing relational business intelligence tools such as Tableau, MicroStrategy, and Qlik. Recently MongoDB is used together with a SQL database creating a data repository also for document storage. MongoDB seems especially suitable to capture real world patient data in clinical trials or collect data from EHRs [11, 12]. For example, a MongoDB implementation enables patients with respiratory disease who could only visit a hospital to measure key data points to capture their data from home [13].

OwnCloud is a cloud-based solution for the storage of confidential data. Because it is cloud-based, it enables users to access data no matter where the data is stored or which device is used to access data. Users are able to decide whether certain data will be transferred to whichever cloud they choose, or whether it will remain within the institution’s own cloud. ownCloud mounts any object store like S3 or OpenStack Swift, and has APIs to integrate with existing tools including SharePoint. ownCloud treats SharePoint as an external storage location, translating ownCloud commands into SharePoint commands enabling mobile, web and sync client access. File access is provided through a single front-end to all of the disparate systems. Two principles characterise this system: a single Point of Access and leave data where it is. In this way, an ownCloud installation can break open data silos by offering a single point of access to files across the organisation or different institutions being able to support a few or up to thousands of storage servers, document libraries, FTP servers and more. Encryption at rest secures the files on the server and still allows sharing among users. With WebDAV, mobile libraries and the ownCloud API as well as several enterprise-only apps allows for secure file sharing ensuring that sensitive data remains under control at all times. Yet it also provides the ease-of-use and mobility needed to streamline healthcare and life science information processes.

ownCloud is also a file sharing solution for healthcare and life science organizations because it combines ease of use with control over sensitive data. Unlike many file sharing systems, which store sensitive data on public cloud servers, ownCloud is integrated into an organization’s IT infrastructure from user directories to security systems. Therefore, ownCloud can be HIPAA compliant. It is possible to set document classification rules and user tags take action to enforce those rules.

The GARR Consortium[3] consists of 58 research hospitals providing the foundation for sharing medical information of all types – including very large MRI images[4]. Based on ownCloud the GARRbox was created for file sharing. The ownCloud platform has created something like the groundwork for an international, open standard of interoperability between different systems of personal cloud storage. In GARRbox the main difference is the so-called "multi-tenancy", which means that each institution has its own storage space for its dedicated data and therefore access is reserved solely to the members of this institution. But GARRbox is able to manage different domains to create multidisciplinary communities for data sharing.

Medical imaging research has to deal with the sharing large and heterogeneous image files. Such challenges lead to the need for secure, federated and functional medical image storage and especially for this area cloud computing promises lower cost, higher scalability and availability. Private cloud storage within an organization based on the ownCloud storage framework was created using the feature of ownCloud to keep images in different file formats and share such images to other researchers in simple way [14]. ownCloud is also a solution used in KKS Leipzig to support clinical trials; files can be uploaded and access-given to the server of the KKS through the private cloud. Such data sharing is used for example for the clinical trial “Effect of daily washing of patients with Octenidine-soaked washing gloves on hospital infections in intensive care units - a randomized, double-blind, cross-over study - EFFECT”[5]. Data for the preparation of the episodes as well as microbiological data are anonymised by the hospitals and, after agreement with the data format, uploaded to ownCloud or sent by post to the KKS. Then the transmitted data is imported into a validated and access-secured study database system [15].

Thus, ownCloud supports an architecture whereby separate private clouds for each institution are established. This approach is used also by Computerome, each Danish hospital has an own private cloud, which enables data sharing between these private clouds through a virtual cloud that exists only for this single purpose[6].

Another example is Hydra employed by the Digital Public Library of America (DPLA), Stanford University and DuraSpace. It is called “next-generation repository solution” that will work for all kinds of institutions, incorporating the capabilities to support networked resources and services in a shared, sustainable, national platform. This platform contains among other features

[3] https://www.garr.it/it/

[4] https://owncloud.com/success-stories/garr/

[5] https://www.viomedo.de/klinische-studien/7081/effekt-taeglichen-waschung-patienten-octenidin-getraenkten-waschhandschuhen-krankenhausinfektionen-intensivstationen-randomisierte-doppel-blinde-cross-over-studie

[6] http://www.computerome.dtu.dk/services1/secure-cloud

Google Scholar metadata tags facilitating indexing and discovery of repository content in a Google Scholar search. It supports IIIF, the International Image Interoperability Framework for publishing and sharing content. It provides an IIIF-compliant Universal Viewer for presenting content to repository user. Another feature is the unique identification of content; on deposit, a unique identifier (Fedora UUID) is generated and assigned to each object. Files are processed on upload to generate key characterization information and other metadata for file management and preservation. In addition, for each stored file a checksum is created that can be verified to ensure the file's integrity is being maintained. Two issues seem to us especially interesting and should be used regardless what solution will be employed by ECRIN. First, the use of Google Scholar metadata tags to allow the search of the repository's content by Google Scholar, without having to rely on an own search engine; second, because each file receives a checksum data integrity is ensured and testable.

Dataverse is a generic repository system that was funded by Harvard University with additional support from the Alfred P. Sloan Foundation, National Science Foundation, National Institutes of Health, Helmsley Charitable Trust, IQSS's Henry A. Murray Research Archive, and many others. Dataverse is an open source web application to share, preserve, cite, explore, and analyse re-search data, with the aim to make data available to others. Researchers, journals, data authors, publishers, data distributors, and affiliated institutions all are enabled to receive academic credit and web visibility.

Each institution can use the Harvard Dataverse software to create a customized Dataverse for researchers and departments to share their research data. In this way the institution becomes a member of the Dataverse repository community consisting of over 27 institutions around the world. World-wide exists 33 installations of the Dataverse software creating 2,872 separate Dataverses with 52,449 Datasets. Only a small number of clinical trials data is stored in Harvard Dataverse, for example: Effect of 5% transdermal lidocaine patches on postoperative analgesia in dogs undergoing Hemilaminectomy or the University of Alberta Libraries' Dataverse, for example: Single-blind, Placebo-challenge Study of Intravenous Fluid Hydration in the Management of Pediatric Migraine in the Emergency Department and Single Incision Device (TVT Secur) Versus Retropubic Tension-free Vaginal Tape Device (TVT) for the Management of Stress Urinary Incontinence in Women: A Randomized Clinical Trial.

Specific repository systems

This group consists of solutions that can be used to create repositories that more specifically address clinical research data. Thus, in contrast to the general repositories that are using file storage, these repositories store data and include data warehouses that store data in tables. Examples are Integrated Data Repository Toolkit (IDRT), i2b2, De-identified Clinical Data Repository (DCDR), European Genome-phenome Archive (Local EGA), Integrating Data for Analysis, Anony-misation and SHaring (iDASH). All these repositories have been stored specific data from humans and clinical studies. For this purpose, data have been converted from the CDMS used in clinical trials to the format of the corresponding data warehouse.

The Open Source software i2b2 provides a tool kit to create a translational research platform for storing biomedical data. The platform contains sophisticated query functions. But, because the platform lacks an easy way for installation, configuration, and the import of source data and its structure is difficult to navigate, the Integrated Data Repository Toolkit (IDRT) was created, which consists of three software additions: first, the i2b2 Wizard automates set-up and administration of i2b2 instances; second, IDRT-Import-Tool is a GUI with an i2b2 server browser and convenient con-figuration wizards to import data into i2b2 from common standard formats such as CSV, CDISC ODM and SQL database; and third, the Ontology-Editor (IOE) allows the editing of i2b2 Ontologies.

One of the use cases for i2b2 is the employment as a Research Data Repository, like the one for the Competence Network for Congenital Heart Defects, which has used i2b2 to integrate it with a clinical trials database. Another use case is DCDR, a data repository containing de-identified data from the clinical systems of the University of Washington Medicine Department[7]. It provides access to pre-programmed datasets in the system that can be obtained at low cost for the cohort. In addition, the ITHS team extracts useful data from electronic medical records to assist with patient cohort identification, trial recruitment, feasibility determination, study design, and more. Recently, Leaf has been developed in partnership with UW Medicine; it is a self-service tool, which provides UW investigators with a user friendly interface for directly querying the electronic health record systems within UW Medicine. The health data in this system is more complete and current than in DCDR. This data is accessed from more than 50 clinical data sources at UW Medicine as well as a network of primary care community clinics.

A secure web-based, graphical query tool to search the database is powered by i2b2. In this case i2b2 is used as a query UI to find eligible patients: researchers can define search criteria and the system returns an aggregate count or summary data of patients who meet these criteria. For example, a query could have the following form: “Provide me a count of patients 18-34 years of age with a diagnosis of diabetes mellitus, who were discharged live within the past six months.” Users of DCDR have to request access by filling out the Access Request Form.

Integrating Data for Analysis, Anonymization and SHaring (iDASH) was founded in 2010 as one of the centers of the initiative National Centers for Biomedical Computing (NCBC) under the NIH Roadmap for Bioinformatics and Computational Biology. It is hosted on the campus of the University of California; San Diego. iDASH collaborates with other NCBCs and develops and disseminates tools. iDASH is an example of a framework for the sharing of sensitive data, which not only consists of a secure data repository, but also of a set of additional tools, for example for the labelling of data, analysis and anonymisation. It deployed cloud storage computing and an

[7] https://www.iths.org/investigators/services/bmi/dcdr/dcdr-dataset-request/

associated policy infrastructure for researchers to share data. iDASH hosts several data sets including different data modalities (whole genomes, transcriptome data, images, specialty reports, clinical trial data, structured and unstructured clinical data) in an annotated data repository, including many related to Kawasaki Disease (one of the world's largest data collections). iDASH's data sharing models include facilitating access to federated databases. A hub for five University of California health systems with 11 million patient's data is included. iDASH tools complement implementations of i2b2 software to be able to count queries with analytical software for privacy-preserving predictive model building. Also policy-based data exchange is enabled by developing a legal framework of data-use agreements (DUA) between data providers and iDASH as data custodian (i.e., honest broker similar to an escrow service), and data recipients and iDASH [16].

Local EGA is a further development of the European Genome-phenome Archive, which has been promoted by EBI for the use of sensitive data, including biomedical and clinical data. The central EGA archive provides a service for permanent archiving and distribution of personally identifiable genetic and phenotypic data, which is the outcome of biomedical research projects. Data stored in the EGA has been collected from individuals whose consent authorise data access only for specific research use to bona fide researchers. Protocols govern how information in EGA is managed, stored and distributed. Local EGAs are local instances of the EGA data repository. In contrast to central EGA, the local EGA repositories store sensitive data locally on the national level. Their metadata are integrated globally with the central EGA. This approach assures that while the local storage stays conform to national requirements and regulations for personal data protection, metadata sharing and storage in the central EGA ensures that all nationally stored data becomes findable. European Genome-phenome Archive contains many clinical studies, mostly GWAS, for example: WTCCC case-control study for Bipolar Disorder, WTCCC case-control study for Inflammatory Bowel Disease, WTCCC case-control study for Hypertension, A Comprehensive Catalogue of Somatic Mutations from a Human Cancer Genome, Genome wide association scan in psoriasis, Breast Cancer Follow-Up Series[8]. The local EGA is divided into several micro-services containing Postgres databases, RabbitMQ message brokers, Inbox SFTP servers (a kind of dropbox), and keyserver (for encryption / decryption keys). Users connect via SSL to the keyserver to initiate re-encryption tasks. The vault moves files from a staging area to the vault storage.

Group of complementary tools and solutions

Generic tools and solutions

This group contains generic solutions of repositories and tools that are employed for all kinds of data. During our search we realised that data repositories are using different additional tools, like for example metadata registries, to deal with data that are part or may become part of a data sharing framework. We therefore saw the need to add this group to the list, although this group is not

[8] https://www.ebi.ac.uk/ega/studies

coherent because it contains completely different services. AWS and Azure are global cloud platforms with lots of different services and options available, Docker is a container management system, and Neo4j is a graph database system. Nonetheless, all of these tools can be used together with a repository system to increase its functionality or security. For example, recently cloud solutions play an increasingly important part in many frameworks for biomedical research. They are added to existing repositories to increase their data sharing capacities, like in the case of EGI and EUDAT data repositories.

The solutions in this group cover: Microsoft Azure, Amazon Web Services (AWS), Orion Metadata Harvester, DataTags, Secure Folder System, Docker, ELIXIR Beacons, Dutch Techcenter for Life Sciences, Open Metadata Registry, Aristotle metadata registry, Open Metadata Repository Services (OMRS), Dataverse, The European Data Portal, Neo4j, and Comprehensive Knowledge Archive Network (CKAN).

Microsoft Azure is a set of cloud services with a focus on business areas including healthcare business. With Azure products it becomes possible to build, manage, and deploy applications as part of a global network. Microsoft developed Azure for healthcare enabling for examole to Control healthcare costs with AI and machine learning.[9] Certain segments of healthcare, for example symptom checking, triage (order of treatment), fraud prevention, precision medicine, can be supported by cloud solutions in combination with AI. Organisations can implement their specialized solution together with solutions of trusted Azure healthcare solution partner, like KenSci Clinical Analytics, Archive 360 (long-term, secure, legal compliant retention and management of data) and CGI ProperPay (reduction of fraud and waste in healthcare claims). In addition, Vivli has developed together with Microsoft an Azure-based platform for cloud access to clinical trials data[10], a cloud platform with data repository, analysis tools and dynamic search engine.

Metadata registries and metadata repositories play an important role for any framework for data storage and data sharing. A metadata registry is a central location in an organization where metadata definitions are stored and maintained in a controlled method; the ISO/IEC 11179 standard sets a norm for metadata registries.[11] A metadata repository is the database where metadata is stored. The metadata repository has additional functionalities compared with registry. It not only stores the metadata items like metadata registry but also adds relationships with related metadata types. Metadata registries will play an important role for data sharing because they are used as central tool to structure metadata and enable interoperability, sharing and re-use of data. In this way, registries act as a central source of schemas or vocabularies for use within a domain[12]. They make data in

[9] https://azure.microsoft.com/en-us/industries/healthcare/usecases/

[10] Healthcare IT News: Vivli to develop Azure-based platform for cloud access to research data (2017), www.healthcareitnews.com/news/vivli-develop-azure-based-platform-cloud-access-research-data

[11] https://en.wikipedia.org/wiki/Metadata_registry; https://en.wikipedia.org/wiki/Metadata_repository

[12] DART: Survey of Metadata Registries (2006): https://www.itee.uq.edu.au/eresearch/projects/dart/outcomes/da3/registries

different repositories searchable. Example of metadata registries in the health domain are: Australian Institute of Health and Welfare - Metadata Online Registry (METeOR) and Cancer Data Standards Repository (caDSR). To enable semantic interoperability for the secondary use of Electronic Health Records (EHRs) for research purposes a federated framework was created to bring together metadata registries and semantic web technologies enabling the integration of data elements through Linked Open Data (LOD) principles, where each Common Data Element (CDE) can be uniquely referenced, queried and processed [17]. Martin Dugas has developed a mapping model between NCI forms in the NCI metadata registry caDSR and the ODM standard [18]. Based on the Medical Data Models portal, an ISO/IEC 11179-compliant metadata registry was implemented to support the re-use of CRF form on the item group and data element levels [19].

In contrast metadata repositories are part of clinical trial data management systems. Using metadata repositories simplifies a standard compliant collection of raw data from all source systems (EDC, labs, ECG, ePRO) in a system-agnostic manner, utilizing study instance metadata specified at study set-up during protocol creation and based on the data standards managed within the metadata repositories. In addition, a metadata repository supports data integration by facilitation of the transformation of data into client-specific or agency-required target data formats. For example, study design and patient data can be mapped into SDTM or other formats through metadata-driven data mapping derived from the metadata repository. Such mapping macros can automate the generation of define.xml for SDTM and annotated eCRFs[13].
The Orion Metadata Harvester created by Orion Governance harvests metadata. Metadata sources are collected from structured data (SQL, ETL, COBOL/JCL), Big Data (Hadoop, HiveDB); and unstructured Data (Email, SharePoint, FileNet). Appliance is a Linux based machine, which runs on a VMWare ESX environment or in a bare metal windows based environment on a VMWorkstation.

DataTags is a suite of tools to help researchers share and use sensitive data in a standardized and responsible way[14]. It is used in some life science applications and it is propagated for sensitive data. The goal of DataTags is to help researchers who are not legal or technical experts to navigate the required knowledge of relevant data privacy laws and make informed decisions when to collect, store, and share private and sensitive data. To our knowledge, DataTags is not used for clinical trials data or healthcare data. But the Institute for Quantitative Social Science is using DataTags for sensitive data to assign security and access requirements to the data when storing social science data in Dataverse[15]. The sharing of sensitive data with DataTags and secure Dataverse is supported by

[13] www.parexel.com/solutions/clinical-research/global-data-operations/clinical-metadata-repository-mdr

[14] https://privacytools.seas.harvard.edu/datatags

[15] https://dataverse.org/presentations/dataverse-cloud-dataverse-and-datatags

the integration with additional tools, like OSF, DataUp, DataBridge, ORCID and iDASH (at Harvard).

Docker represents a containerization format, which is a type of operating system-level virtualization that has become an open standard in the industry. Containerization offers the potential to package software systems, like the i2b2 platform components, into standalone executable packages that are agnostic to the underlying host operating system. Docker provides a way to run applications securely isolated in a container, packaged with all its dependencies and libraries. It is based on the idea that packaging existing apps into containers immediately improves security, reduces costs, and gain cloud portability. Docker packaged applications including their dependencies are put together into an isolated container, which makes them portable to any infrastructure. For translational medicine, the i2b2 platform was implemented as Docker containers. Three Docker container images: WildFly, database, and web, were developed to encapsulate the three major deployment components of i2b2. These containers isolate the core functionalities of the i2b2 platform, and work together to provide its functionalities [20]. An architecture was developed for genomics analysis in a clinical setting using Galaxy and Docker [21], and OpenClinica the open source software for Electronic Data Capture (EDC) for clinical research has been dockerised by TMFe.V. [22]; their container uses the official Apache Tomcat Docker image (tomcat 7) and OpenClinica[16].

ELIXIR Beacons provide discovery services for genomic data in the European Genome-phenome Archive (EGA) and in ELIXIR Nodes, using the Beacon technology developed by the Global Alliance for Genomics and Health (GA4GH)[17]. The GA4GH Beacon is an open data sharing platform that allows genomic data centres to make its data discoverable. Users can use following queries: “Do any of these data resources have genomes with this allele at that position?” The generated search query result informs a researcher as to whether making a data access request is required for their research, saving valuable time and resource. In principle, the beacon method is about searching large pools of similarly structured data at their location. It was suggested by the coordinator of CORBEL as solution for the search of clinical trials data. One should consider this option to create a beacon for searching at different clinical trials databases at ECRIN members in different countries in a legal and secure way. Because ELIXIR Beacons are discovery services for genomic data in EGAs and ELIXIR Nodes, it would be necessary to structure the clinical trials data in a different way and to collaborate with ELIXIR. For example, in 2017 ELIXIR has extended its collaboration with GA4GH to implement the Beacon technology across more ELIXIR Nodes, now including ELIXIR Belgium, EMBL-EBI (UK), ELIXIR Spain, ELIXIR Switzerland, ELIXIR Netherlands, ELIXIR Finland, ELIXIR Sweden, and ELIXIR France. It was declared that ELIXIR Nodes can handle sensitive, personal data through their secure archives and their comprehensive, end-to-end solutions for data privacy that go beyond simple download protection. In this context, one could install an EGA at a partner institution of ECRIN to store ECRIN clinical trials data and

[16] https://hub.docker.com/r/tmfev/openclinica/

[17] https://www.elixir-europe.org/about-us/implementation-studies/beacons

this EGA could be connected by Beacons with other ELIXIR nodes to run searches[18] [19]. In the area of research on rare diseases, ELIXIR is already providing support services by building a portal through which authorised researchers can access rare disease data from repositories and catalogues around Europe. Preliminary work is being done in an Implementation Study addressing the visualization of aligned genomics data for rare diseases (RD-Connect) as a driver for real-time access of controlled data at the EGA[20].

Open Metadata Registry is available under an open licence. It provides means to identify, declare and publish through registration metadata schemas, schemes (controlled vocabularies) and application profiles. In addition, Open Metadata Registry supports machine mapping of relationships among terms and concepts in those schemes (semantic mappings) and schemas (crosswalks). Thus this solution could be a nice addition to a clinical trials data repository by unifying different metadata schema in use and by integrating standardised vocabularies and data catalogues.

Aristotle metadata registry is another open-source metadata registry compliant with the requirements of ISO/IEC 11179 specification. It represents a new way to manage and federate content built on and extending the principles of leading metadata registries. The MDR is built on Django web framework and the 11179 standard, allowing institutions to easily run their own metadata registries.

Open Metadata Repository Services (OMRS) enables the integration of metadata that is distributed amongst a number of metadata repositories either through a call interface which is provided by an OMRS connector, or via linked data URLs that enable a metadata entity to have a relationship with a metadata entity in a different repository. In this way, the Open Metadata Repository Services (OMRS) enable metadata repositories to exchange metadata. In principle, it could be used to build an exchange service between databases that store clinical trials data.

It is a general problem for researchers in the life sciences that they lack a place to easily assess datasets from different providers in terms of services provided and metadata richness. The software YummyData [22] addresses this problem by providing based on periodically polls a curated list of SPARQL endpoints, monitoring the states of their Linked Data implementations and content. In this way, YummyData[21] can improve the findability and reusability of life science datasets provided as Linked Data and to foster its adoption. Such a tool should also include the search for clinical trials

[18] https://wikis.bris.ac.uk/download/.../Senf-ELIXIR.pdf?

[19] ERA-Net for Research Programmes on Rare Diseases: http://www.erare.eu/Infrastructures/elixir

[20] https://www.elixir-europe.org/use-cases/rare-diseases

[21] Database URL: http://yummydata.org/

resources so that researchers can be provided with the actual states of the many repositories that contain clinical trials data.

The European Data Portal harvests the metadata of public sector information available on public data portals across many European countries. Information regarding the provision of data and the benefits of re-using data is also included, together with information about datasets, catalogues, metadata quality, licensing assistant, SPARQL manager, and statistics tools[22]. Data are free for use and reuse for commercial or non-commercial purposes. It also aims to help foster the transparency and the accountability of the institutions and other bodies of the EU. The EU Open Data Portal is managed by the Publications Office of the European Union. It contains information about the clinical trial domain like the European Union Clinical Trials Register, EnprEMA Network Database (EnprEMA), and European Ombudsman - Annual Reports. It contains clinical research data sets about topics like: Youth attitudes on drugs; Price, purity and potency of drugs in Europe; Death rate due to chronic diseases by sex and for example, 5 data sets about Diabetes research (Self-reported screening of cardiovascular diseases and diabetes risks by sex, age and degree of urbanisation, …). The metadata catalogue can be searched via an interactive search engine (Data tab) and through SPARQL queries. The portal is built with Drupal content management system and CKAN, the data catalogue software was developed by the Open Knowledge Foundation. It uses Virtuoso as an RDF database and has a SPARQL endpoint [22]. Its metadata catalogue is built on international standards such as Dublin Core, the data catalogue vocabulary DCAT and the asset description metadata schema ADMS.

Neo4j is a graph platform. It takes a connections-first approach to data by using persisting relationships and connections through every transition of data from logical model, to implementation in a physical model, to operation using a query language. The foundation of representing connected data is known as a graph. Neo4j's Graph Platform is built around the Neo4j native graph database. It supports transactional applications and graph analytics, data integration expedites distilling tabular data and big data into graphs, Cypher graph query language is the bridge to big data analytic tooling, and graph visualization and discovery. Recently Neo4j is used in the Life Sciences and Healthcare, because it makes it easier to work with highly-connected information [23]. Managing, storing and querying connected information is the focus of a graph database like Neo4j. Neo4j is used for research in meta-proteomics and cancer research[23]; for the querying of disease networks [24]; it is used for connecting protein databases in a large graph model and for creating "Reactome" databases of human protein interaction pathways. The company Zephyr Health is using graph databases for the collection and storage of data from about 3500 sources of clinical research data, including public sources (ClinicalTrials.gov and PubMed), as well as private data from partner organisations and pharma industry to help pharmaceuticals and medical device companies to understand for example, who publishes the most and the best research in a certain

[22] https://en.wikipedia.org/wiki/EU_Open_Data_Portal

[23] Using Graph Technology to Fight Cancer: https://neo4j.com/news/using-graph-technology-to-fight-cancer/

area[24]. In addition, the company is integrating diverse healthcare data by using MongoDB and Neo4j to connect patients to healthcare treatments and therapies[25].

Comprehensive Knowledge Archive Network (CKAN) is a framework, a powerful data management system that makes data accessible by providing tools to streamline publishing, sharing, finding and using data[26]. It is software for building a catalogue and repository for datasets. The system can store datasets, or simply hold metadata for datasets hosted externally. It provides discovery services such as keyword and map-based searching, and can harvest metadata from and syndicate its own metadata to other repositories. CKAN is most easily installed on a server running Ubuntu 12.04 LTS, for which pre-compiled packages are available. It may be installed on many other platforms, since it supports Python, PostgreSQL, Apache SolR, Jetty, and Java. Filters have been written for harvesting metadata from other CKAN instances, CSW servers, WAFs (Web Accessible Folders), ArcGIS portals, Geoportal Servers, Z39.50 databases, DCAT (Data Catalogue Vocabulary) RDF/XML sources, and other repositories. A CKAN instance can syndicate its metadata through CSW, RSS/Atom feeds, RDF (XML, N3), CSV and JSON dumps. CKAN's codebase is maintained by Open Knowledge International. The system is used both as a public platform on Datahub and in various government data catalogues, such as the UK's data.gov.uk, the Dutch National Data Register, the United States government's Data.gov and the Australian government's "Gov 2.0"[27]. . CKAN is used by governments, organisations and communities.

Specific tools and solutions

This group consists of tools that were specifically created to deal with clinical data and other sensitive data for translational research, including MOLGENIS, TranSMART, TraIT, TSD (Tjenester for Sensitive Data), TRYGGVE, UK Data Service Secure Lab, DataSHIELD, Electronic Health Records for Clinical Research (EHR4CR), Aircloak, The Secure Data Vault (SDV), Jisc DataVault, European Genome-phenome Archive, ELIXIR Use Case for human data research, Computerome, CREDEN-TIAL, Rare Disease (RD) Connect, The European Network for Cancer Research in Children and Adolescents (ENCCA), European Medical Information Framework

[24] Kempe S: Graph Databases have Impact on Healthcare Sector (2014): http://www.dataversity.net/graph-databases-impact-healthcare-sector/

[25] Integrating Diverse Healthcare Data using MongoDB and Neo4j: https://neo4j.com/blog/healthcare-data-mongodb-neo4j/

[26] http://www.dcc.ac.uk/resources/external/ckan

[27] https://en.wikipedia.org/wiki/CKAN

(EMIF), Open Source Registry System for Rare Diseases in the EU (OSSE), Janus Data Repository, FHIR, Medical Data Space and ePouta IaaS Cloud. Several tools, like MOLGENIS, TranSMART, TraIT, TSD (Tjenester for Sensitive Data), TRYGGVE, UK Data Service Secure Lab, DataSHIELD and Computerome have experience with the storage and sharing of clinical trials data.

Some tools don't deal with the storage of sensitive data, but with only one aspect of sensitive data sharing. Such tools were considered, because they may become part of a framework for sensitive data. This group includes: Heterogeneous Proxy Re-Encryption (H-PRE), Framework for Sensitive Data Sharing and Privacy Preserving on Big-Data, MONOMI, Privacy-Sensitive Sharing Framework, and A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains.

MOLGENIS is a modular web application to support molecular genetics research; but it is also used for biobanking, research of rare diseases, and creating patient registries. MOLGENIS provides researchers with a user friendly software to capture, exchange, and use large amounts of data. Use Cases for MOLGENIS comprise the Finnish disease database (FINDIS), Animal observation database (AnimalDB), eXtensible Genotype and Phenotype database (XGAP), Design of Genetical Genomics Experiments (designGG), Mouse Resource Browser (MRB), Nordic GWAS control database and the Bacterial microarrays database (MOLGEN-IS).

TranSMART is a knowledge management and content analysis platform for translational research. It enables analysis of integrated data for the purposes of hypothesis generation, hypothesis validation, and cohort discovery. In this way, tranSMART bridges basic science with clinical practice data by merging multiple types of data from disparate sources into a common analysis environment. For this purpose it uses Amazon Cloud. tranSMART data sources contain clinical trials data, which are stored in the tranSMART data warehouse, which can be searched by using data mining techniques. tranSMART includes many type of sensitive data, like demographics (e.g. age, sex and ethnicity), physical examinations, patient history, medical diagnoses, medical treatments, laboratory test results (e.g. standard blood test or advanced biomolecular test), pathology reports in free text, radiology images, and clinical outcomes (e.g. survival rates). Phenotypic data are stored using the i2b2 data model consisting of an entity attribute value pair schema. OpenClinica can be linked to tranSMART; a converter transfers ODM data into the tabular format of tranSMART. Though, tranSMART can also use SDTM format. But this is also the weak point of tranSMART, the necessity to convert data into a special tabular format to make them searchable. Use cases of tranSMART are the Transforming Research and Clinical Knowledge in Traumatic Brain Injury (TBI), UK Severe Asthma Registry, Sage Bionetworks Collaboration (providing the comparator arm of a J&J clinical trial for CTCAP (Computed tomography of the chest, abdomen, and pelvis), and the IMI (Innovative Medicines Initiative) project eTRIKS (European Translational Information & Knowledge Management Services). The Research and Clinical Knowledge in Traumatic Brain Injury (TBI) is a 5-year NINDS funded study to determine the relationships among clinical, neuroimaging, cognitive, genetic and proteomic biomarkers with the aim to validate biomarkers and outcome measures for clinical trials.

TraIT project[28] started as an initiative from the Center for Translational Molecular Medicine (CTMM). TraIT enables integration and querying of information across four domains of translational research: clinical, imaging, biobanking and omics data. The TraIT project combines different tools, including tranSMART, Molgenis, for imaging XNAT, ImageHub, Keosys, for integration with cBioPortal, for pathology tEPIS, and for biosamples data Molgenis catalogue. TraIT enables integration and querying of information across the four major domains of translational research: clinical, imaging, biobanking and experimental (any-omics) with a particular focus on the needs of muticenter projects. Rather than developing new software, TraIT uses proven solutions that can be adopted or adapted to the specific needs of translational research projects. TraIT already supports over 475 research projects spanning almost 4000 individual researchers.

In the domain of clinical research data, TraIT has the objective to make data management processes more efficient, improve data quality and enable re-use of clinical research data[29]. As data capture tool to collect clinical research data and metadata OpenClinica is in use since 2011. The most important reasons for selecting OpenClinica are that it is Open Source license, is web based, supports all types of clinical studies, and built on leading, independent standards. Ldot was selected as a scheduling tool to support research personnel in the planning of clinical studies and logistic support functionalities.

TSD (Tjenester for Sensitive Data) is a platform to collect, store, analyse and share sensitive data in compliance with Norwegian regulation regarding necessary data privacy protection. TSD is used by researchers working at University of Oslo, but as part of the EOSC-Hub its outreach is increased to European researchers. TSD is primarily an IT-platform for research, even though in some cases it is used for clinical research and even commercial research. In principle, TSD provides a secure centralized vault where data and backup are stored. Inside a secure environment reside all TSD services, databases and a data management infrastructure. Researchers working within a project, access the project-dedicated resources in TSD through remote connection. A large set of software is installed in TSD to allow analysis of the data inside the secure environment. Examples of clinical projects in TSD are Cognitive Behavioural Therapy and Dental Fear (CT.gov), The Nordic Atrial Fibrillation and Stroke Study (NOR-FIB), the Atrial Fibrillation in Cryptogenic Stroke and TIA Study protocol. Usage of TSD contains: 1 TB storage, 1x Windows server with 4CPUs and 8GB RAM, 1x Linux RHEL6 virtual machine with 2CPUs and 4GB RAM, backup of 1 TB storage area, standard software.

Fast Healthcare Interoperability Resources (FHIR) is a new standard for exchanging health care information electronically.[30] It is based on previous standards including Health Level Seven v2 and v3, the Reference Information Model, and Clinical Document Architecture; it specifies a RESTful

[28] http://www.ctmm-trait.nl

[29] http://www.ctmm-trait.nl/trait-domains/work-package-1-clinical-research

[30] Fast Healthcare Interoperability Resources: Draft Standards for Trial Use 2. 2015. https://www.hl7.org/fhir/2015May/index.html. Accessed December 15, 2015.

API to access resources. An interface was developed to bring patient data from i2b2 repositories into the Fast Healthcare Interoperability Resources (FHIR) format, referred to as a SMART-on-FHIR cell [25]. The cell serves FHIR resources on a per-patient basis, and is implemented as an i2b2 server plug-in, consisting of modules for authentication, REST, i2b2-to-FHIR converter, resource enrichment, query engine, and cache.

FHIR is beginning to play an important role for establishing interoperability between EHR and EDC systems to streamline clinical investigations[31], especially in the generation and use of Real World Evidence (RWE)[32], and to support the use of eSource to pre-populate CDASH Case Report Forms using a CDISC ODM API[33]. The clinical data management system REDCap was enabled to use FHIR based data sources [27].

DataSHIELD is an infrastructure and set of R packages that enables the remote and non-disclosive analysis of sensitive research data. Users are not required to have prior knowledge of R, but should be able to use the R console. It is appropriate for the analysis of harmonised individual level data at multiple locations. In this federated framework each data location installs the server-side DataSHIELD infrastructure that holds a snapshot of the harmonised data to be analysed. One of the locations also installs and manages the DataSHIELD client portal, which is the mechanism by which users are authenticated to send analysis commands within the DataSHIELD infrastructure. This infrastructure can also be used to analyse individual level data held at only one location. In this case, the data server-side DataSHIELD infrastructure is installed in addition to the DataSHIELD client portal.

Aircloak is one example of a high-security vault system that allows for immediate, safe and legal sharing of sensitive data. The Aircloak software is located between databases and the outside analysts. It acts like a filter for personal data enabling much higher fidelity analytics than existing anonymisation or data masking solution. It can work with SQL and NoSQL databases on the user side. Aircloak's sits between the existing "primary-use" database and the untrusted "secondary use" analysts and the applications; then it filters queries and answers to ensure user anonymity.

The Danish National Life Science Supercomputing Centre called Computerome is a HPC Facility specialized for Life Science demands. Users include research groups from all Danish Universities and large international research consortiums as well as users from industry and the public Health Care Sector. They all can benefit from the fast, flexible and secure infrastructure and the ability to combine different types of sensitive data to perform analysis. Computerome is physically installed at the DTU Risø campus. It is the official supercomputer of ELIXIR Denmark, a member of ELIXIR. Services cover the separate storage with HIPAA-compliant auditing, storage

[31] http://wiki.hl7.org/index.php?title=201801_Clinical_Research

[32] https://www.cdisc.org/sites/default/files/resource/Use_of_Fast_Healthcare_Interoperability_Resources_in_the_Generation_of_Real_World_Evidence.pdf

[33] https://www.lexjansen.com/phuse-us/2018/tt/TT16.pdf

of private copies of reference databases like TCGA and Ensemble, Batch Queuing system, and provision of a wide selection of preconfigured and regularly updated scientific tools. Computerome is characterised by its extreme scalability (from 1 CPU VM to 1000+ CPU), its complete separation from the internet or the use of monitored and filtered one way traffic, access to specialised services e.g. GPU nodes, Data Analysis and Visualization services, Dynamic Pipelines, cloud bursting and Data collection to analysis service.

The Medical Data Space is a joint initiative of the Fraunhofer groups ICT Technology and Life Sciences and another example for a secure area and vault framework. The Medical Data Space (MedDS) is a virtual space that supports secure exchange and integration of medical and health-related data from diverse sources, using standards and shared governance models with the aim to improve the quality of diagnostics, preventive and therapeutic measures. MedDS protects the digital sovereignty of the data owner (the patient, hospital, physician, drug company, etc.).

Heterogeneous Proxy Re-Encryption (H-PRE) is a special encryption method. When data owners store their data as plaintext in clouds, they lose the security of their cloud data due to arbitrary accessibility, specially accessed by the un-trusted cloud. In order to protect the confidentiality of data owner’s cloud data, a new method to be used is the encryption of data by the data owner before storing the data in the cloud. However, the employment of traditional encryption algorithms cannot solve the problem, since it is hard for data owners to manage the process and their private keys, if they want to securely share their cloud data with others in a fine-grained manner. Heterogeneous proxy re-encryption (FH-PRE) system can be used to protect the confidentiality of cloud data. By applying the FH-PRE system in the cloud, the data of the data owner can be securely stored in cloud and shared in a fine-grained manner. It even provides the secure data sharing between two heterogeneous cloud systems, which may be equipped with different cryptographic techniques.

MONOMI is a system for securely executing analytical processes over sensitive data on an un-trusted database server. MONOMI works by encrypting the entire database and running queries over the encrypted data. It employs methods like split client / server query execution, which can execute arbitrarily complex queries over encrypted data, as well as several techniques to improve the performance for the analysis, including per-row precomputation, space-efficient encryption, grouped homomorphic addition, and pre-filtering. A prototype of MONOMI is running on top of Postgres[34].

Discussion

The result of the survey of possible solutions for the storage of individual participant data from clinical trials is that there is no ideal solution available; there is no solution that satisfies all stakeholder requirements and at the same time is Open Source. Though, tranSMART and Molgenis

[34] https://github.com/stephentu/monomi-optimizer

are coming near, they are very focused on translational medicine requirements. It seems to us that there is a general move in research IT away from monolithic solutions to ones that can add components and tools according to the demands of the research done.

Our search showed that many solutions already used to store data in institutions allow for some kind of restricted access, but that there was no single solution that met all the requirements for storing clinical trials data. Clinical trials data repositories must be part of a larger system that encompasses different tools for encryption, anonymisation, authentication & authorisation, cloud solution, a secure access area, metadata repositories, an identifier system and tools for search and analytics.

In summary, the best solution for the storage of clinical trials data is to use an existing repository if it is compliant with the CORBEL requirements or to build an ecosystem for sensitive data usage by positioning a general data repository system, like DSpace, in the center of additional tools and extensions, like data management solutions (e.g. CRIS), encryption, anonymization, connections to secure analysis areas, secure vault systems and cloud storage systems, as well as to establish connections to metadata registries and open data providers, like figshare.

In order to gain technical experience and to come forward with the decision to develop a new repository for clinical trial data or to use / adapt an existing one, it was decided to create a pilot repository, using a specific repository system that is in widespread use. From the assessment done so far, DSpace was identified as one of the promising candidates. Currently, a pilot installation is under development.

One reason for DSpace was that it is used by very different service providers, like libraries, institutions and even by organisations that share research data. Duraspace[35] listed seven types of DSpace usage scenarios: DSpace as 1 Institutional Repository Platform, 2 Digital Collection Management, 3 Current Research Information System (CRIS), 4 Data Repository, 5 Learning Object Repository (LOR), 6 Digital Preservation System and 7 Web Content Management System (WCMS). Such broadness of deployment shows that the system has to be flexible and scaleable.

For example, Edinburgh DataShare[36] installed DSpace for their clinical trials data. And for CERN it turned out that their Invenio database was not sufficient and had to be complemented by DSpace[37]. We therefore decided to begin with the evaluation of DSpace for its usability as a basic data repository for the pilot. In fact, DSpace is the underlying repository for many institutions and museums to store and share research data (World Bank Open Knowledge Repository, Digital Access to Scholarship at Harvard, DSpace at MIT, BORA at Bergen University College, research library at CERN, etc.). Features that make DSpace strong in this area are the integrated support for

[35] https://wiki.duraspace.org/display/DSPACE/DSpace+Positioning

[36] Edinburgh DataShare - A DSpace Data Repository: https://www.era.lib.ed.ac.uk/handle/1842/3201

[37] http://webzine.web.cern.ch/webzine/9/papers/3/index.html

Dublin Core metadata, customizable workflows, submission forms that can be defined on a per collection basis, and support of OAI-PMH, possibilities to define embargo periods, access control features and a very good indexing by Google Scholar. Therefore it will be possible to search clinical trials in our pilot by using google.

With DSpace one can join a global network of similar archives or repositories at many research institutions. This makes it for us easy to share metadata with Edinburgh DataShare. It is the most widely used repository software platform with more than 2,000 installations worldwide. For example, DSpace at Cineca is a non-profit consortium in Italy made up of 70 Italian universities, four national research centres, and the Ministry of Universities and Research. In this way Cineca resembles the ECRIN network, which is also a non-profit organisation consisting of 10 member or observer countries with national ECRIN offices, 11 certified data centres and numerous clinical trial units (CTUs) in the member countries.

DSpace has several features that promise an easy and cost-free installation. DSpace is an out-of-the-box Open Source software package for creating repositories focused on delivering digital content to end users and providing a full set of tools for managing and preserving content within the application (Fig. 1).

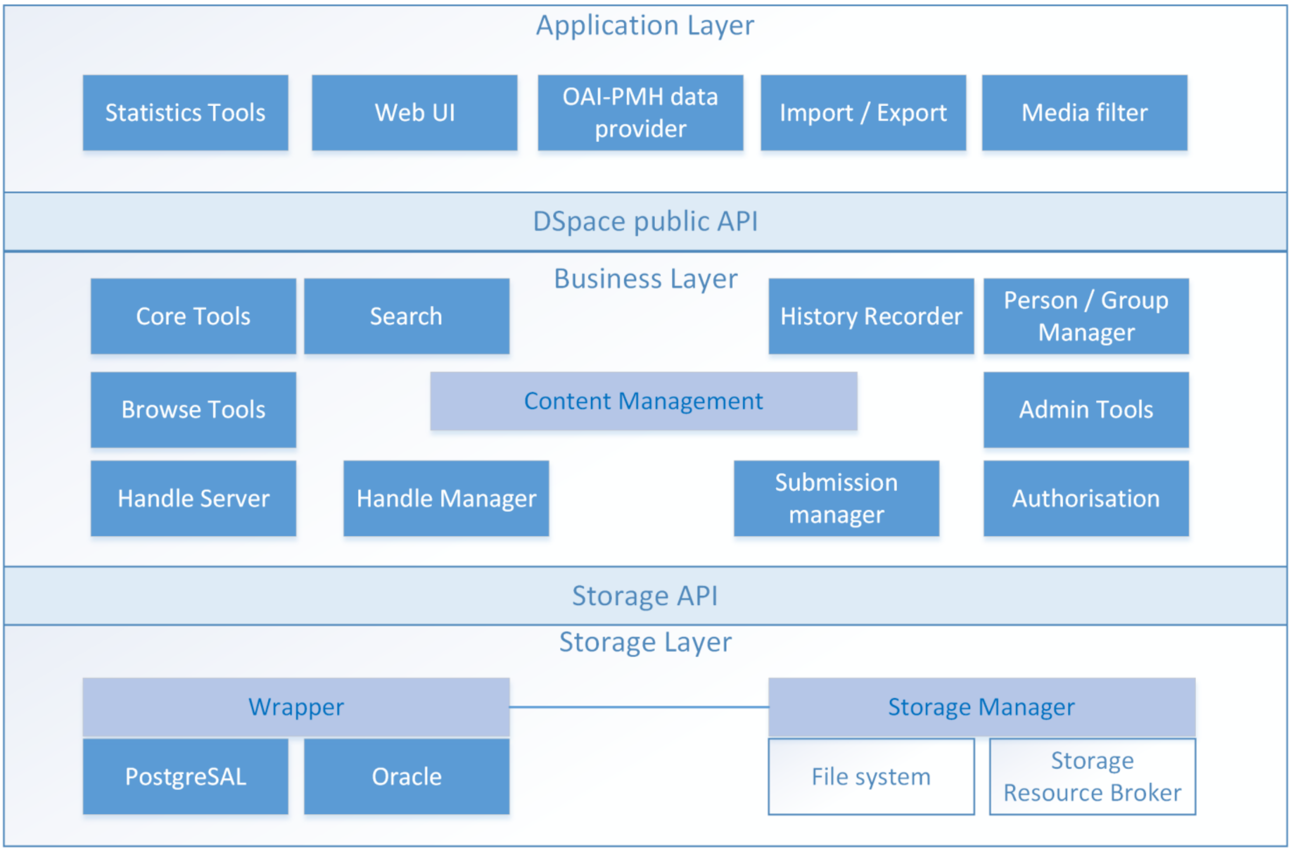

Fig. 1: The different components of the DSpace data repository architecture (modified from [28])

For example, the famous Dryad Digital Repository is a curated resource that makes data of underlying scientific publications discoverable and freely reusable. Dryad promotes an ecosystem, where research data is openly available, integrated with publications, and routinely re-used. Dryad is built upon DSpace[38] and all its customizations that are not available within the main DSpace distribution are available from a code repository under an open source license.

There exist several extension possibilities for DSpace. Apollo repository is the DSpace based database of the University of Cambridge Repository. It is used by the School of Clinical Medicine to store research data as well as, scholarly works and theses in the area of medicine. DSpace-CRIS is a free open-source extension of DSpace especially created for the management of research data. IRIS, Institutional Research Information System, is the new CRIS solution developed by Cineca, offering ORCID enhanced interoperability for a national system. In contrast to DSpace, the overall concept of a CRIS system is broader and encompasses rich objects for staff, projects, grants etc. to achieve data visibility. Nevertheless, several institutions have successfully implemented DSpace as a CRIS component, like Hong Kong University. Additional features that make DSpace strong as scientific data repository are persistent URLs and unique identifiers, item and bitstream versioning, checksum generation and verification and the bitstream format registry. DSpace can also be used as Digital Preservation System, which is a system to safeguard assets for the long term. Features that make DSpace strong as Digital Preservation System are built-in checksum checker, bitstream format validator, distributed asset storage, AIP import and export and linking with Duracloud. Anyhow, DSpace does not provide a vault system for highly secure, hacker proof storage.

Recently many institutions in Germany have installed DSpace; the German user group alone has 25 member institutions. During the German DSpace User Group Meeting in Berlin[39] institutions presented their DSpace solution. Often DSpace is part of a research management system to store and distribute research documents, publications, theses and research data. Such an implementation is often financed by a DFG project for 5 years (3 years implementation, 2 years configuring and testing) amounting to about 1 Mio EURO of funding. This seems to be the minimal amount necessary for such a project. Maintenance costs are normally not considered, because maintenance efforts are minimal and done by the normal library employees. Nonetheless, end of this year the new version of DSpace 7 will be available, which will introduce several changes and a completely new user interface. An update to DSpace 7 will be a considerable effort that will costs many additional person months. Because most funding projects will have ended, the question arises who can afford an update to the newest version.

Our pilot installation for individual level clinical trials data doesn’t need to provide a research management addition. It should only store the data and give access to authorised users. Thus there is need to install CKAN or CRIS on top of DSpace. But what would be useful would be the addition of services specifically to deal with sensitive data, which are tools for anonymization, encryption and perhaps a vault system. During the German DSpace User Group Meeting the case of clinical

[38] The repository: Technology: http://datadryad.org/pages/repository

[39] https://wiki.duraspace.org/display/DSPACE/DSpace+Anwendertreffen+2018

trials data in DSpace was discussed[40]. It was stressed during the meeting that DSpace is a solution for open access and not suited for clinical trials data. On the other hand, increasingly DSpace base repositories are asked to store medical data and sensitive data of the social sciences. The problem seems to be that as long the data in DSpace is accessible through the internet, a sufficient security cannot be guaranteed. To achieve this level of security data in DSpace must be encrypted or a vault system should be added, this could be a DSpace instance on a PC isolated from the internet, but connected to the DSpace instance that is connected to the internet.

The basic information structure of DSpace is quite simple: a single tree of community, collection and items.

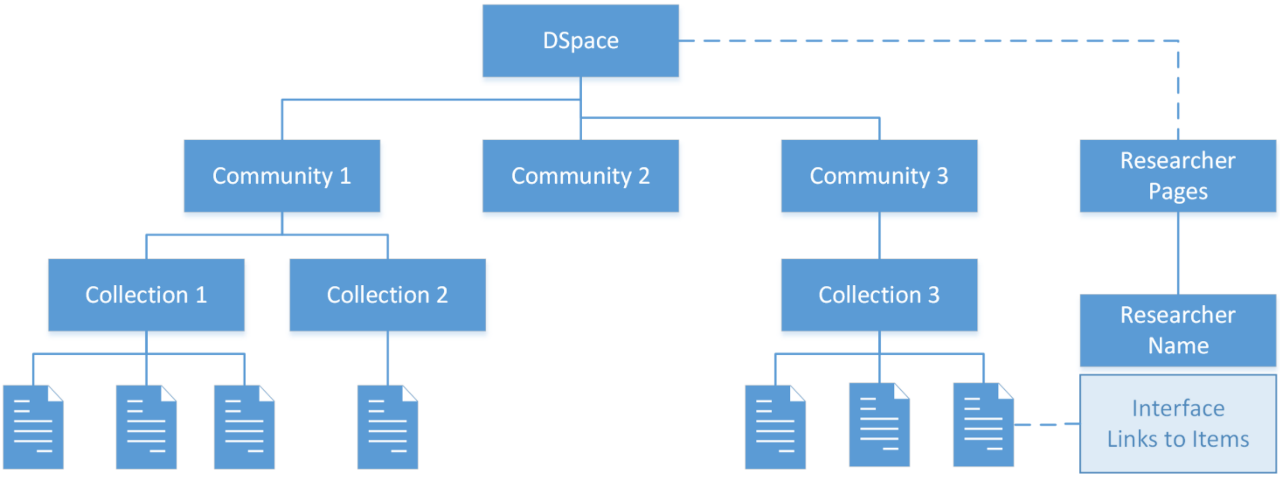

Fig. 2: Basic information architecture of DSpace (modified from [28]

In this structure, for the use of ECRIN clinical trials data, the community could be an ECRIN member state or an institution, or a research network; the collection represents a single clinical trial container for all documents, data and metadata of a clinical trial; at the lowest level the items represent the different data objects, like study protocol, statistical analysis plan, clinical trials database, publications, outcome tables, etc. The items can represent the corresponding document of a link to the document.

In addition, the standard ingest workflow of DSpace has to be only minimally modified for clinical trials data. In the workflow the data provider sends a request for uploading, is being authorised, and uploads the data. A DSpace manager, normally library stuff, reviews the uploaded content and gives an approval and the data can be uploaded and stored in DSpace. For the uploaded content descriptive metadata (standard is Dublin Core) and a licence has to be provided. It is already possible to store the data in a restricted access folder so that the data is not openly accessible. Only the metadata is openly readable. The problem with security for sensitive data is that though data is not accessible, the data stays hackable, until it is transferred to a vault separated from the internet.

[40] Thanks to Pascal-Nicolas Becker (The Library Code GmbH) to put DSpace for clinical trials data on the agenda for discussion.

But most institutions using DSpace for sensitive data seem to accept this risk, or are not aware of it and store sensitive data in a restricted access folder system without any additional protection. The additional costs for adding an encryption system or a vault to DSpace would be several person months. Several additional modifications for adapting DSpace to the secure usage of sensitive data derived from our survey.

Modifications of DSpace and estimated efforts in person months:

1. Modification of metadata schema, extension of the basic Dublin Core: 1 PMs
2. Integration of anonymization / encryption service: 5 PMs
3. Addition to a vault system (an isolated DSpace instance): 2 PMs
4. Connection to TSD or DataSHIELD: 5 PMs

Extension with the possibility for Heterogeneous Proxy Re-Encryption would make it possible to analyse data on the encrypted clinical trials database. Other useful extension would be the linking to a system for efficient authentication and authorization, like the ELIXIR Authentication and Authorization Infrastructure. Such a system should also consider the rights for data reuse given by the consent of study participants.

References

[1] Ohmann C, Banzi R, Canham S, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. BMJ Open. 2017;7(12):e018647. doi:10.1136/bmjopen-2017-018647.

[2] Ohmann C, Canham S, Banzi R, Kuchinke W, Battaglia S. Classification of processes involved in sharing individual participant data from clinical trials. F1000Research. 2018;7:138. doi:10.12688/f1000research.13789.2.

[3] Tudur Smith C, Hopkins C, Sydes MR, et al. How should individual participant data (IPD) from publicly funded clinical trials be shared? BMC Medicine. 2015;13:298. doi:10.1186/s12916-015-0532-z.

[4] Chervitz SA, Deutsch EW, Field D, et al. Data Standards for Omics Data: The Basis of Data Sharing and Reuse. Methods in molecular biology (Clifton, NJ). 2011;719:31-69. doi:10.1007/978-1-61779-027-0_2.

[5] Hopkins C, Sydes M, Murray G, et al. UK publicly funded Clinical Trials Units supported a controlled access approach to share individual participant data but highlighted concerns. Journal of Clinical Epidemiology. 2016;70:17-25. doi:10.1016/j.jclinepi.2015.07.002.

[6] Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington (DC): National Academies Press (US); 2015 Apr 20. 3, The Roles and Responsibilities of Stakeholders in the Sharing of Clinical Trial Data. Available from: https://www.ncbi.nlm.nih.gov/books/NBK286000/

[7] CDISC: Fostering Responsible Data Sharing through Standards (2014), available: https://www.cdisc.org/fostering-responsible-data-sharing-through-standards

[8] So D, Knoppers BM (2017) Ethics approval in applications for open-access clinical trial data: An analysis of researcher statements to clinicalstudydatarequest.com. PLoS ONE 12(9): e0184491. https://doi.org/10.1371/journal.pone.0184491

[9] Sensitive Data Session (Beyond EUDAT 2020), Sept 26-27, 2017, available: https://eudat.eu/sensitive-data-session-beyond-eudat-2020

[10] https://www.researchgate.net/publication/310424834_EUDAT_Working_Group_on_Sensitive_Data

[11] Gomez A, Eijo, Juan E, Martínez von Scheidt M Baum A, Luna D, Quirós F. (2013). MongoDB: An open source alternative for HL7-CDA clinical documents management. 10.13140/RG.2.1.3033.7128.

[12] Freire SM, Teodoro D, Wei-Kleiner F, Sundvall E, Karlsson D, Lambrix P. Comparing the Performance of NoSQL Approaches for Managing Archetype-Based Electronic Health Record Data. Carter KW, ed. PLoS ONE. 2016;11(3):e0150069. doi:10.1371/journal.pone.0150069.

[13] Hosford T, Deverell B. Enabling Clinical Research in the Real World with Sensors, Node.js and MongoDB. Presentation at MongoDB World 2017: Session Recordings. Available:

https://explore.mongodb.com/vidyard-all-players/mongodb-world-presentations-crystal-b-tom-hosford-brenda-deverell-6-20-2017

[14] Hani AFM, Paputungan IV, Hassan MF, Asirvadam VS. and Daharus M. Development of private cloud storage for medical image research data, 2014 International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, 2014, pp. 1-6. doi: 10.1109/ICCOINS.2014.6868433

[15] Prüfplan: Effekt der täglichen Waschung von Patienten mit Octenidin-getränkten Waschhandschuhen auf Krankenhausinfektionen in Intensivstationen – eine randomisierte, doppel-blinde, Cross-Over Studie – EFFECT. 08.11.2016 Status der Fassung: final2.0.

[16] integrating Data for Analysis, Anonymization and Sharing: https://commonfund.nih.gov/sites/default/files/Integrate_Data_for_Analysis_Anonymization_and_Sharing_%28IDASH%29_at_the_University_of_California_San_Diego.pdf

[17] Sinaci AA, Gokce B, Erturkmen L. A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. Journal of Biomedical Informatics. 2013; 46, (5): 784-794. doi.org/10.1016/j.jbi.2013.05.009

[18] Interoperability in Clinical Research: From Metadata Registries to Semantically Annotated CDISC ODM. Authors Philipp Bruland, Bernhard Breil, Fleur Fritz, Martin Dugas Pages564 - 568DOI10.3233/978-1-61499-101-4-564SeriesStudies in Health Technology and Informatics, Ebook, Volume 180: Quality of Life through Quality of Information

[19] Dugas M. Design of case report forms based on a public metadata registry: re-use of data elements to improve compatibility of data. Trials 2016; 17:566-571. doi.org/10.1186/s13063-016-1691-8

[20] Wagholikar KB, Dessai P, Sanz J, Mendis ME, Bell DS, Murphy SN. Implementation of informatics for integrating biology and the bedside (i2b2) platform as Docker containers. BMC Medical Informatics and Decision Making. 2018;18:66. doi:10.1186/s12911-018-0646-2

[21] Digan W, Countouris H, Barritault M, et al. An architecture for genomics analysis in a clinical setting using Galaxy and Docker. GigaScience. 2017;6(11):1-9. doi:10.1093/gigascience/gix099

[22] Yamamoto Y, Yamaguchi A, Splendiani A. YummyData: providing high-quality open life science data. Database; 2018, 1 January 2018, bay022, doi.org/10.1093/database/bay022

[23] European Public Sector Information Platform. Understanding the European Data portal. ePSIplatform Topic Report No. 2016/03, February 2016. https://www.europeandataportal.eu/sites/default/files/2016_understanding_the_european_data_portal.pdf

[24] Yoon B-H, Kim S-K, Kim S-Y. Use of Graph Database for the Integration of Heterogeneous Biological Data. Genomics & Informatics. 2017;15(1):19-27. doi:10.5808/GI.2017.15.1.19.

[25] Park Y, Shankar M, Park B and Ghosh J. Graph databases for large-scale healthcare systems: A framework for efficient data management and data services. 2014; IEEE 30th International Conference on Data Engineering Workshops, Chicago, IL, 2014: 12-19. doi: 10.1109/ICDEW.2014.6818295

[26] Lysenko A, Roznovăţ IA, Saqi M, Mazein A, Rawlings CJ, Auffray C. Representing and querying disease networks using graph databases. BioData Mining. 2016;9:23. doi:10.1186/s13040-016-0102-8.

[27] Wagholikar KB, Mandel JC, Klann JG, Wattanasin N, Mendis M, Chute CG, Mandl KD, Murphy SN. SMART-on-FHIR implemented over i2b2, Journal of the American Medical Informatics Association. 2017; Volume 24(2):398–402, doi.org/10.1093/jamia/ocw079

[27] Leroux H, Lawley M, Gibson S. REDCap under FHIR: Enhancing electronic data capture with FHIR capability. 2017; In: Health Informatics Conference; August 6-9 2017; Brisbane, QLD, Australia. IOS Press Series; 2017. 2.

[28] The DSpace Developer Team: DSpace 6.x Documentation. 27 June 2018. https://wiki.duraspace.org/display/DSDOC6x

Annex 1: Stakeholder requirements

No	Requirement
1	Provision of individual-participant data (IPD) should be promoted, incentivised and resourced
2	Clinical trial datasets should be considered legitimate and citable products of research.
3	Clinical trial datasets should have an associated persistent and globally recognised identifier.
4	Gaining consent to secondary use of data should become a standard procedure.
5	Boards overseeing the data sharing process should be established, ideally at the level of data repository
6	The right to request access to data should not be limited to specific professions or roles.
7	The results and methodology of data analysis and the trial documents should be publicly available and therefore be deposited in an appropriate repository.
8	The processing of data sharing access requests should be explicit, reproducible, and transparent
9	This processing should minimise the additional bureaucratic burden on all parties concerned.
10	Repositories should be encouraged to make the interface presented to secondary users as consistent as possible.
11	Metadata repositories should be developed, sustained and connected
12	A metadata schema suitable for describing all repository data objects linked to clinical trials should be developed and implemented
13	Mechanisms should be developed to make it easy to assign unique identifiers to all datasets and documents that are made available for data sharing.
14	The harvested metadata should be imported into a collection of 'metadata repositories' for clinical research data objects
15	A single point of entry for users and associated search facilities should be provided.
16	Tools should be developed to help data generators to complete metadata fields in the generic schema
17	Tools should be developed to enable the regular harvesting of metadata from repositories
18	Services to support de-identification of datasets should be established.

19 Shared data should be structured, described and formatted using widely recognised data and metadata standards.

20 Datasets should be made available for sharing in one or more standardised file formats that can be read by a wide variety of different systems.

21 Repositories with clinical research data objects should use a generic schema for its data objects, or a schema that can be easily mapped to it

22 Data sharing models should be based on the concept of data 'stewardship' rather than data 'ownership'.

23 Services to support and store documents should be provided.

24 Time for making IPD data and documents available for re-use should be monitored.

25 Explicit consent for data sharing should be provided at the same time as the informed consent for the clinical trial participation.

26 Consent for the secondary use of IPD should be as broad as possible.

27 Consent should clarify the reasons for data sharing, and the general benefits of data sharing in clinical research.

28 Trial participants should have the right to withdraw specifically their consent for data sharing.

29 Among the various data standards the CDISC standards should be used for defining and coding data and metadata in a consistent way.

30 A standard based 'Data Use Agreement (DUA)', which specifies conditions for data access and re-use, should be implemented.

31 Access to IPD shall be accompanied by a statement of compliance with basic rules designed to promote the fair sharing of data.

32 Not only the IPD datasets, but additional clinical trial data objects should be made available for sharing (e.g. protocols, clinical study reports, statistical analysis plans, blank consent forms)

33 Data and trial documents made available for sharing should be transferred to a suitable data repository

34 The data objects should be properly prepared, stored f securely or long terms, and be subject to rigorous data governance.

35 Repositories for clinical data and data objects should be compliant with defined quality criteria.

36 Ingest of any data objects to repositories should be subject to a formal agreement that defines roles, rights and responsibilities of the data generators and the repository managers.

37 Before data is shared, it shall be de-identified by removing identifiers to minimize the risk of re-identification.

38 Shared data should remain pseudonymous unless that is not allowed by the relevant legislation

39 Additional information that may allow re-identification should be stored securely and not be shared.

40 Standard procedures and techniques for de-identification should be applied, and be fully documented

41 An assessment of risk for re-identification of trial participants in de-identified datasets should be performed.

42 Any re-identification of data subjects shall be forbidden.

43 Clinical trial datasets should always be associated with metadata that describes characteristics of each data item (e.g. type, code, name, possibly an ontology reference), as well as metadata that describe schedule and design of the trial.

44 Access to IPD and trial documents should be as open as possible and as closed as necessary, to protect participant privacy and reduce the risk of data misuse.

45 A range of different access types to shared data and documents including different forms of controlled access should be established.

46 Mechanisms to collect and display user feedback about the processes of accessing data and data sharing should be developed and implemented by repositories or by third parties.

47 The analysis environment should allow different datasets from different host repositories to be combined on a temporary basis.

48 Any dataset or document made available for sharing should be associated with concise, publicly available and consistently structured metadata to ease discovery of datasets, describing not just the data object itself, but also, how it can be accessed.

49 The generic metadata scheme should include a common identifier scheme for clinical research data objects.

Annex 2: List of Examples of IT Services/Tools for Sensitive Data

1. Group: Generic Repository Systems

1. Name of service / tool	**DSpace**
Contact address / person, if available	Duraspace group
Webpage of the tool	http://www.dspace.org/ **Release version**: 6.0, http://www.dspace.org/latest-release **Documentation:** https://wiki.duraspace.org/display/DSDOC/
Country it is used	Most of the counties in academic and government
Cross-country use?	Yes
Short description of the tool	**Description**: DSpace is an out-of-the-box open source software package for creating repositories focused on delivering digital content to end users and providing a full set of tools for managing and preserving content within the application. DSpace is the most widely used repository software platform (open source or proprietary), with more than 2,000 installations worldwide representing a continuously growing and active user community. **History**: DSpace was originally developed by MIT Libraries and Hewlett-Packard (HP) Labs. Since its initial open source release in 2002, the platform has been guided by a global community of committers, developers, repository managers, and other stakeholders who contribute to project governance. DSpace became a DuraSpace project in 2009 when the Fedora Commons and DSpace organizations merged to form DuraSpace. **Cost**: Open source software, no charge. DSpace is distributed under the terms of the BSD open source license.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	**Technical Aspects** **Operating System:** Written in Java, tested under Linux, Windows, and Mac OSX **License:** BSD **Other prerequisite software:** Java 7 or 8, Apache Maven, Apache Ant, Relational Database (PostreSQL or Oracle), Servlet 3.0 container (Tomcat 7+ or similar).
Modules / architecture / components included	**Key Features** **Application Architecture:** DSpace is a full stack web application, consisting of a database, storage manager and front end web interface. The architecture includes a specific data model with configurable metadata schemas, workflows and browse/search functionality.

	Modern, RESTful Web UI (coming soon): DSpace 7.0 will feature a completely rewritten web user interface based on the Angular 2 javascript platform. **Built-in workflows:** Originally designed for libraries, the embedded DSpace data model and approval workflows are familiar to librarians and archivists. **Built-in search engine:** DSpace comes packaged with Apache Solr, an open source enterprise search platform that enables filtered (faceted) searching and browsing of all objects. The full text of common file formats is searchable, along with all metadata fields. Browse by interfaces are also configurable. **Unlimited File types:** DSpace can store any type of file. In addition, it auto-recognizes files of most common formats (e.g., DOC, PDF, XLS, PPT, JPEG, MPEG, TIFF). **Metadata:** By default, DSpace uses a Qualified Dublin Core (QDC) based metadata schema. Institutions can extend that base schema or add custom QDC-like schemas. DSpace can import or export metadata from other major metadata schemas such as MARC or MODS. **Tools/plugins:** DSpace comes with a suite of tools (batch ingest, batch export, batch metadata editing, etc.) and plugins for translating content into DSpace objects. Additionally, commercial plugins are available through service providers. **Security:** DSpace provides its own built-in authentication / authorization system, but can also integrate with existing authentication systems such as LDAP or Shibboleth. **Permissions:** DSpace allows you to control read/write permissions site-wide, per community, per collection, per item and per file. You may also delegate administrative permissions per community or per collection. **Disaster Recovery:** DSpace allows you to export all of your system content as AIP (Archival Information Packages) backup files. These AIPs can be used to restore your entire site, or restore individual communities, collections or items. **OAI-PMH / SWORD (v1 and v2) / OpenAIRE / Driver:** DSpace complies with standard protocols and best practices for access, ingest, and export. **REST:** DSpace provides RESTful APIs in accordance with modern web standards. **Configurable Database:** Organizations can choose either PostgreSQL or Oracle for the database in which DSpace manages items and metadata. **Configurable File Storage:** Files in DSpace can be stored either using a local filesystem (default) or a cloud-based solution, such as Amazon S3.

	Data Integrity: On upload, DSpace calculates and stores a checksum for each file. Optionally, you may ask DSpace to verify those checksums to validate file integrity. **Languages:** DSpace is available in over 20 languages.
What data is stored in the tool	DSpace preserves and enables easy and open access to all types of digital **content including text, images, moving images, mpegs and data sets.** And with an ever-growing community of developers, committed to continuously expanding and improving the software, each DSpace installation benefits from the next.
Research use cases / projects / studies the tool is used (including collaborations)	The most common use of the DSpace software is by **academic and research libraries** as an **open access repository** for managing their faculty and student output. There are also many organizations using the software to host and manage subject based, dataset or media-based repositories. The following provides the major use case categories and several examples of each. **Institutional Repository:** Deep Blue is the University of Michigan's permanent, safe, and accessible service for representing their rich intellectual community. Its primary goal is to provide access to the works that make Michigan a leader in research, teaching, and creativity. Deep Blue uses DSpace as a publications' repository for published objects including journals. Click here to learn more about their repository service. **Image Repository:** Rice University’s Travelers in the Middle East Archive, TIMEA. TIMEA is a digital archive that focuses on Western interactions with the Middle East, particularly travels to Egypt during the nineteenth and early twentieth centuries. TIMEA offers electronic texts such as travel guides, museum catalogs, and travel narratives, photographic and hand-drawn images of Egypt, historical maps, and interactive GIS (Geographic Information Systems) maps of Egypt and Cyprus. **Audio / Video Repository:** Georgia Tech SMARTech, or Scholarly Materials And Research @ Georgia Tech, is a repository for the capture of the intellectual output of the Institute in support of its teaching and research missions. SMARTech connects stockpiles of digital materials currently in existence throughout campus to create a cohesive, useful, sustainable repository available to Georgia Tech and the world. In addition to hosting the 4th International Conference on Open Repositories (OR09), Georgia Tech also created a video repository of the recordings from the DSpace User Group Meeting held in conjunction with the conference. **Museum / Cultural Heritage:** New York University’s Afghanistan Digital Library retrieves and restores materials published in Afghanistan that are in danger of disappearing. The objective of the project was to collect, catalogue, digitize and provide access to as many publications as possible in an effort to reconstruct an essential part of Afghanistan's cultural heritage. "Afghan scholars are enormously excited about The Afghanistan Digital Library as this is material that don't have access to currently," says Dr. Michael Stoller, the project

	manager of ADL and Director of Collection and Research Services at New York University. **Government Records / Reports:** Policy Archive is a comprehensive digital library of public policy research containing over 24,000 documents. It is a digital archive of global, non-partisan public policy research that makes use of the Internet technology to collect and disseminate summaries and full texts, videos, reports, briefs, and multimedia material of think tank, university, government, and foundation-funded policy research. It offers a subject index, an internal search engine, useful abstracts, email notifications of newly added research, and will expand to offer information on researchers and funders, and even user-generated publication reviews. It will also over time, grow to include policy content from international and corporate organizations. **Subject:** The Belgian Poison Centre created a repository to store information on the composition and toxicity of various products, medicines, plants, animals and fungi. The site also offers guidance to healthcare professionals and a list of the most frequent inquiries of the Center. The Center fields more than 50,000 calls per year and is subsidized by the Federal Public Service of Public Health. **Learning Resources:** Ontario Council of University Libraries Cooperative Online Repository for Information Literacy (CORIL) is an initiative to support information literacy instruction among Canadian universities. CORIL is a repository for both open collection or the peer reviewed materials. **Federated Repositories / Networked Instances:** The Texas Digital Library is a multi-university consortium dedicated to providing the digital infrastructure to support a fully online scholarly community for institutions of higher education in Texas. Formed in 2005 by four Texas members of the Association of Research Libraries, the TDL has extended membership in the consortium to any of the state’s institutions of higher learning. Through the establishment of shared policies and standards, forums for professional interaction, expertise in digital collections and preservation, and robust technical services, the TDL aims to increase the availability of the enormous intellectual capital of Texas universities and to preserve it for future generations.
Comments	

2. Name of service / tool	**Samvera**
Contact address / person, if available	The Hydra Project, Samvera's original name, was founded in 2008 by: Stanford University, University of Virginia, University of Hull, Fedora (now part of DuraSpace)
Webpage of the tool	http://samvera.org/
Country it is used	Most of the countries in Europe and North America
Cross-country use?	Yes
Short description of the tool	SAMVERA IS AN **OPEN SOURCE REPOSITORY FRAMEWORK** Samvera software was conceived as an open source repository framework. That is to say that we set out to create a series of free-to-use software "building blocks" that could put together in various combinations to achieve the repository system that an institution needed – as opposed to building a "one size fits all" solution. The framework as it exists today consists of a number of Ruby gems that can be combined, configured and adapted to serve a wide variety of needs, as you can see on our "applications and demos" page. Some of our adopters started with just the basic building blocks, but then a more common approach became to find another institution whose use case was similar, clone their Samvera variant and then adapt it to more closely fit local needs.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	**Extensible** The Samvera framework is architected to extend features based on specific concerns and use-cases. Established features with long-term community commitment are available, such as: • Browse Everything allows your rails application to access user files from cloud storage, such as Dropbox, Skydrive, Google Drive, Box, and a server-side directory share • LDP provides a Linked Data Platform and allows publication of RDF data
Modules / architecture / components included	**Three bundles of Samvera framework**, which are currently in various stages of development: **Avalon** – a time-based media solution **Hyrax** – is a Ruby gem that includes much of Samvera's functionality. It is the basis on which users can build their own, customized version of Samvera. **Hyku** – the product from the "Hydra-in-a-Box" project. The Digital Public Library of America (DPLA), Stanford University and DuraSpace have partnered to extend the existing Samvera codebase to

	build, bundle, and promote a feature-rich, robust, flexible digital repository that is easy to install, configure, and maintain. Hyku can be installed locally or run in the cloud and is based on Hyrax. A number of service providers, including DuraSpace themeselves, are – or will soon be – offering cloud-based, hosted versions.
What data is stored in the tool	Within Samvera framework, one body, **the digital repository**, supports multiple heads or tailored, **content type-specific applications**.
Research use cases / projects / studies the tool is used (including collaborations)	**We could not find medical usage.** **Boston Public Library** As the oldest public library system in the country, the Boston Public Library (BPL) holds a vast wealth of culturally significant materials in its collections, which they have been digitizing for many years. **Northwestern University** Northwestern University Library (NUL) is home to a large number of distinctive, rare, and unique collections including the largest separate collection of Africana materials outside of the continent and the largest transportation collection in the world **Notre Dame University** By Rick Johnson, Hesburgh Libraries, University of Notre Dame In early 2009, as we approached the challenge of creating a digital repository and related services for our archival collections, we saw a problem much larger than ourselves. After an earlier failed attempt, it was clear there was no vended solution that was mature and flexible enough to meet our needs. We were also just beginning to learn what it meant to run a digital repository, and the prospect of creating and maintaining solutions by ourselves was extremely daunting.
Comments	All Samvera's software is **free and open source**, available under an Apache 2 license.

3. Name of service / tool	**Hydra-In-a-Box**
Contact address / person, if available	Contact us: hybox-contact@googlegroups.com
Webpage of the tool	http://hydrainabox.samvera.org/
Country it is used	
Cross-country use?	Yes
Short description of the tool	The Digital Public Library of America (DPLA), Stanford University and DuraSpace are partnering to extend the **existing Samvera** community (formerly "Hydra project") codebase and its vibrant and growing community to build, bundle, and promote a feature-rich, robust, flexible digital repository that is easy to install, configure, and maintain. This next-generation repository solution -- "Hydra-in-a-Box" -- will work for institutions large and small, incorporating the capabilities and affordances to support networked resources and services in a shared, sustainable, national platform. The overall intent is to develop a digital collections platform that is not just "on the web," but "of the web."
Type of activity the tool supports (project, service, collaboration, platform, etc.)	**Descriptive Metadata** Hyku supports a set of 17 metadata elements for describing collections and works. **Faceted search and browse** – Powered by Blacklight, an intuitive interface makes it easy to search and find content quickly. **Google Scholar metadata tags** – Facilitates indexing and discovery of repository works in a Google Scholar search. **Supports IIIF** – International Image Interoperability Framework for publishing and sharing content, and provides the IIIF-compliant Universal Viewer for presenting content to repository users. **Manage citations** – Exportable to EndNote.
Modules / architecture / components included	**Unique identification** – On deposit, a unique identifier (Fedora UUID) is generated and assigned to each object. **File characterization** – File are processed on upload to generate key characterization information and other technical metadata necessary for ongoing file management and preservation. **File audit –** The checksum of a stored file can be verified to ensure the file's integrity has been maintained. **ResourceSync** – Implementation of this emerging standard protocol for facilitating the synchronization of online content, useful search engine optimization and for feeding content to aggregators like DPLA.

	IIIF – Implementations of the Image API and Presentation API make repository content interoperable with other IIIF-compliant systems and applications. **RESTful HTTP API –** Basic, general purpose interface for enabling programmatic interoperability.
What data is stored in the tool	**Content Types** Content of any variety is supported, and any file format can be uploaded. Hyku currently has two Work types available. **Generic Work** – Hyku's original model for any piece of content or "work". Appropriate for any file type. Includes generic, broadly applicable descriptors and provides a menu of common Resource Type values which can be applied to characterize the nature of the content. **Image Work** – The first distinct content type developed in Hyku. Similar to the Generic Work, with additional, optional descriptors particularly relevant to works featuring image-based content, such as photographs or illustrations, as distinct from text-based content, time-based media, dataset, etc. Plans to add support for enhanced sequencing and labeling files within an image work. **Content types for future development**: • Research dataset • Software • Thesis / dissertation • Oral history • Book • Newspaper
Research use cases / projects / studies the tool is used (including collaborations)	We could not find medical usage. We are continually working to develop Hyrax to support more diverse use cases, so that over time the Hyku repository application will be capable of wearing a number of hats on its "head" – an institutional repository, a data repository, a digital collections management system, an audiovisual repository, etc.
Comments	

4. Name of service / tool	**Fedora**
Contact address / person, if available	Mailing address: DuraSpace 9450 SW Gemini Drive #79059 Beaverton, OR 97008 Telephone: (607) 216-4548 Fax: (607) 697-0418 Email: info@fedora-commons.org
Webpage of the tool	http://fedorarepository.org/
Country it is used	Over 400 organizations in more than 35 countries have registered their Fedora installations.
Cross-country use?	Yes
Short description of the tool	Fedora is a robust, modular, open source repository system for the management and dissemination of digital content. It is especially suited for digital libraries and archives, both for access and preservation. It is also used to provide specialized access to very large and complex digital collections of historic and cultural materials as well as scientific data. The Fedora project is led by the Fedora Leadership Group and is under the stewardship of the DuraSpace not-for-profit organization providing leadership and innovation for open source technology projects and solutions that focus on durable, persistent access to digital data. In partnership with stakeholder community members DuraSpace has put together global, strategic collaborations to sustain Fedora which is used by more than three hundred institutions.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	**Native Linked Data Support:** Fedora is a Linked Data Platform server. It speaks RDF by default and participates as a native citizen of the read/write web. **Open Data, Open Formats:** It's just as easy to get your data out of Fedora as it is to get your data in - export your data in an open RDF format that isn't dependant on Fedora.
Modules / architecture / components included	**Features** **Powered by Fedora:** Fedora is a key component of larger application frameworks such as Samvera and Islandora. **Standards Based Services:** Fedora provides a set of core repository services via RESTful APIs using modern web standards. **Any File, Any Size:** Fedora can store, preserve, and provide access to any type of file - no restrictions! **Extensible Architecture:** Fedora is designed to integrate with other applications and services to provide search, discovery, and more.

	Advanced Storage Options: Fedora provides a variety of storage options for your files and metadata, including file systems, databases, and more. **Pluggable Security:** Secure your digital assets with pluggable authorization modules: role-based, XACML, or Web Access Control. **Preservation Ready:** Fedora provides key preservation services, such as fixity checking, an audit trail, versioning, and import/export. **Highly Scalable:** Fedora can handle millions of files and metadata records. **RESTful APIs:** Fedora provides core services via well-documented RESTful APIs that clients can rely on. **Powerful Extensions:** Fedora is more than a set of core services - plug-in modules provide OAI-PMH dissemination, SWORD deposit, and more! **Message Based Workflows:** Use the built-in messaging framework to build powerful, scalable workflows. **Easy Deployment:** Fedora deploys easily as a WAR file into any servlet container. **Batch Operations:** Bundle a series of actions together into a single repository event to achieve better consistency and performance.
What data is stored in the tool	Digital libraries and archives, very large and complex digital collections of historic and cultural materials as well as scientific data.
Research use cases / projects / studies the tool is used (including collaborations)	We could not find medical usage. Fedora has a worldwide installed user base that includes academic and cultural heritage organizations, universities, research institutions, university libraries, national libraries, and government agencies.
Comments	

5. Name of service / tool	**Clinical Data Repository**
Contact person	Email: ctsi@umn.edu Phone: 612-625-CTSI (2874)
Webpage	https://www.ctsi.umn.edu/researcher-resources/clinical-data-repository https://research.medicine.umich.edu/office-research/institutional-review-boards-irbmed/guidance/repository-overview
Country	University of Minnesota, USA
Cross-country	
Short description (only a few sentences)	Researchers have access to data in a clinical data repository that houses the electronic medical records of more than 2 million patients.
Type of activity (project, service, collaboration, platform, etc.)	
Modules/components included	
Data included	The data in the University of Minnesota’s Clinical Data Repository comes from the electronic health records (EHRs) of more than 2 million patients seen at 8 hospitals and more than 40 clinics. Data is available for hospital visits starting in 2011; for Fairview Health Services clinic visits from 2005; and for University of Minnesota Physicians clinic visits starting from 2011. All start dates are approximate and depend on the date for the adoption of Epic at each particular site. For each patient, data is available regarding the patient's demographics (age, gender, language, etc.), medical history, problem list, allergies, immunizations, outpatient vitals, diagnoses, procedures, medications, lab tests, visit locations, providers, provider specialties, and more.
Research use cases (including collaborations)	The data in this repository can be used for biomedical research, including recruitment planning, retrospective cohort studies, and observational studies
Comments	

6. Name of service / tool	**MongoDB**
Contact address / person, if available	Berlin WeWork Potsdamer Platz Stresemannstraße 123, 10963 Berlin Germany
Webpage of the tool	https://www.mongodb.com/
Country, it is developed and used	MongoDB is a global company. MongoDB was founded in 2007 by Dwight Merriman, Eliot Horowitz and Kevin Ryan – the team behind DoubleClick. DoubleClick now **owned by Google**.
Cross-country use?	Yes. MongoDB has more than 4,900 customers in over 85 countries. Headquartered in New York, with offices across North America, Europe, and Asia-Pacific
Short description of the tool	MongoDB is a document database with the scalability and flexibility that you want with the querying and indexing that you need.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	**Client Libraries (Drivers)** for C, C++, C# and .NET, Java, Node.js, Perl, PHP, Python, Ruby, Scala **MongoDB Stitch** is a backend as a service that provides an HTTP API to MongoDB, integration with other services. **The MongoDB Connector for Business Intelligence (BI)** allows users to create queries with SQL and visualize, graph, and report on their MongoDB Enterprise data using existing relational business intelligence tools such as Tableau, MicroStrategy, and Qlik. **MongoDB Connector for Spark** provides integration between MongoDB and Apache Spark Libraries for use with MongoDB datasets: Datasets for analysis with SQL (benefiting from automatic schema inference), streaming, machine learning, and graph APIs. **Integrated Feature Set.** Analytics and data visualization, text and geospatial search, graph processing, event-driven streaming data pipelines, in-memory performance and global replication allow you to deliver a wide variety of real-time applications on one technology, reliably and securely. MongoDB stores data in flexible, JSON-like documents, meaning fields can vary from document to document and data structure can be changed over time. The document model maps to the objects in your application code, making data easy to work with. Ad hoc queries, indexing, and real time aggregation provide powerful ways to access and analyze your data
Modules / architecture / components included	**MongoDB Atlas** is a cloud service for running, monitoring, and maintaining MongoDB deployments, including the provisioning of dedicated servers for the MongoDB instances. In addition, Atlas

	provides the ability to introspect collections, query backups, and migrate data from existing MongoDB replica set into an Atlas cluster. **MongoDB Cloud Manager** provides a complete package for managing MongoDB deployments. **MongoDB Ops Manager** is a package for managing MongoDB deployments. Ops Manager provides Ops Manager Monitoring and Ops Manager Backup, which helps users optimize clusters and mitigate operational risk. **MongoDB Compass** is designed to allow users to easily analyze and understand the contents of their data collections within MongoDB and perform queries, without requiring knowledge of MongoDB **Distributed Data Platform.** MongoDB can be run within and across geographically distributed data centers and cloud regions, providing new levels of availability and scalability.
What data is stored in the tool	**Flexible Data Model.** NoSQL databases emerged to address the requirements for the data we see dominating modern applications. Whether document, graph, key-value, or wide-column, all of them offer a flexible data model, making it easy to store and combine data of any structure and allow dynamic modification of the schema without downtime or performance impact.
Research use cases / projects / studies the tool is used (including collaborations)	**Healthcare Solutions:** Using MongoDB, healthcare providers can create a single application that provides **a 360-degree view of the patient**, aggregating patient, doctor, procedure and other types of information in a single data store. **Zephyr Health** is integrating diverse healthcare data by using two databases: MongoDB and Neo4j connecting patients to the healthcare treatments and therapies they most need. MongoDB is used as document store, which holds all profile information for each doctor. **Enabling Clinical Research in the Real World with Sensors**, Node.js & MongoDB at Koneksa Health (New York). Koneksa uses Linux on Amazon Web Services, including a number of Amazon services such as CloudWatch for logging, EC2 for servers and more. It pairs MongoDB as primary data store containing all application data and a summarized form of the data, with Amazon Simple to store the raw data. With a tracker watch, for instance, the raw data is stored in AWS' Simple Storage Service, then the application will transform and normalize the data into a JSON format used by Mongo and ultimately used by scientists for analysis. At the application layer, it uses Node.js and Mongoose for object modeling. Everything is in JavaScript, from the data in the API to the database. The modeling Respiratory Data with MongoDB using seamless JS integration and embedded document structure. Participant's measurements can be queried with $lookup (Example: finding all measurements for a user). **Population Management for At-Risk** Demographics. Certain populations are known to be prone to certain diseases. For instance, men over a certain age are more likely to have specific types of cancer.

	With MongoDB’s flexible data model, providers of lab testing, genomics and clinical pathology can ingest, **store and analyze a variety of data types** from numerous sources all in a single data store **Platforms and Services:** Amazon EC2, Red Hat Enterprise Linux, dotCloud, Rackspace Cloud, Red Hat OpenShift, VMware Cloud, Foundry, Microsoft Azure, Windows Quick Links and Reference Center **Other Cases:** Storing Log Data, Pre-Aggregated Reports (MMAPv1), Hierarchical Aggregation, Product Catalog, Inventory Management, Category Hierarchy, Metadata and Asset Management, Storing Comments
Comments	**MongoDB** is free and open-source, published under the GNU Affero General Public License

7. Name of service / tool	**ownCloud**
Contact address / person, if available	Contact: Telefon: +49 911 14888690 Telefax: +49 911 56981566 E-Mail: info@owncloud.com Holger Dyroff Rathsbergstr. 17 90411 Nürnberg
Webpage of the tool	https://owncloud.com/overview/ https://owncloud.org/
Country it is used	
Cross-country use?	Yes
Short description of the tool	Organizations that must share confidential data internally and externally rely on ownCloud - the open platform for better productivity and security within digital collaboration. It enables users to access data no matter where it is stored or which device is used. Users are able to decide whether certain data will be transferred to whichever cloud they choose, or whether it will remain within the enterprise's own On-Premises Cloud.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	**Existing Infrastructure:** ownCloud mounts any object store like S3 or OpenStack Swift, and has APIs to integrate easily with existing tools and the corporate infrastructure including: **SharePoint:** ownCloud treats SharePoint as an external storage location, translating ownCloud commands into SharePoint commands and enabling mobile, web and sync client access. With SharePoint integration, ownCloud users can directly access their SharePoint document libraries. **Windows Network Drives:** ownCloud administrators can integrate one or more network drives within a single ownCloud instance, treating them as external storage and allowing users to access, sync and share files stored on the Windows network drives. **S3, Swift, Ceph, FTP:** ownCloud leverages storage that already exists, including S3, Swift, Ceph or FTP.
Modules / architecture / components included	**Security Controls** **Admin Set User/File-level Permissions:** User or file-level permissions can be defined when and where files are shared. Access expiration dates and restrictions can be set at multiple levels. Plus, administrators can use File Firewall to create rules that control access to ownCloud servers based on user connections, time intervals, geographic locations and more. **File Firewall:** The File Firewall provides a policy engine for the ownCloud instance, prohibiting access to files that do not meet

	standards. Rules and operators (AND, OR, NOT, EQUAL) are configured by the admin based on attributes of a request. And the results of each rule evaluation can be logged for reporting. A File Firewall rule can also evaluate a tag on a file and determine access based on the specific file request. **Encryption:** ownCloud provides two levels of Encryption capabilities; encrypting server data at rest, and supporting encryption for data in motion. Another option for customers who want encryption if motion is for them to use SSL.Additionally, ownCloud gives customers the ability to manage their key stores and as well as access/manage the reading and writing of files. Admins choose and implement the key manager of their choice (theirs, ours or a different one altogether) or replace the AES-256 cipher with something different like a cipher of your choosing. ownCloud is the only vendor to provide this capability. Our Encryption 2.0 is built modularly with the ability to swap out components. Encryption from ownCloud is delivered as an app that is easily and quickly integrated with your existing infrastructure. **Key management / choose algorithm:** By default, our competitors manage encryption keys in the cloud which exposes them to the same vulnerabilities as the cloud. ownCloud allows you to manage keys in your enterprise key store. You may also create your own key manager, and write an app to use your own encryption solution. **File Integrity Checking:** To prevent file corruption the integrity of up- and downloaded files is automatically verified by comparing their unique checksums before and after transfer. **Authentication(SSO / SAML 2.0):** Single Sign On (SSO) is supported and Shibboleth, a SAML-based authentication, is integrated with ownCloud’s web-front end, ownCloud mobile apps, and desktop clients. As users are managed by those services, ownCloud automatically acquires and implements the associated authentication. **AD/ LDAP:** Built-in wizards allow IT to integrate ownCloud with Active Directory or LDAP or customers may choose custom authentication mechanisms as needed for their environment. **2-Factor Authentication:** Integrated 2-Factor Authentication Provides More Security. The authentication method allows that additional technologies and tokens can to be used via plugins. This not only improves access security, but also provides administrators with an option for disabling individual tokens. Time-based one-time passwords (TOTP) enable users to automatically increase the security of their accounts by using services like Google Authenticator or the open-source implementation of the TOTP standard. **Virus Scan:** When enabled, by default, uploaded files are scanned with ClamAV, preventing the potential for automated distribution of infected files or integrated with external virus scanners. **Auditability / Logging:** Not only does ownCloud allow IT to control each user’s permissions, but it also enables a full audit trail—allowing IT to understand how, when and where data is accessed and shared. A single app allows admins to log account level activities such as logins

	to ownCloud as well as what users do with files on the server. This provides admins the basic information they need for compliance reporting and auditing of ownCloud usage and the tools to actively follow file sharing activities. The use of a SIEM solution like Splunk or other log readers is reported.
What data is stored in the tool	Any type of data
Research use cases / projects / studies the tool is used (including collaborations)	It is used for **medical research image data sharing**. ownCloud is a file sharing solution for **healthcare** and life science organizations, because it combines ease of use without compromising control over sensitive data. Unlike consumer-grade file sharing services, which store sensitive data on public cloud servers, ownCloud is deployed within an organization’s IT infrastructure. It is designed from the ground up to integrate with existing directory, security and rights management systems, ensuring that data governance policies are enforced end-to-end. ownCloud’s flexible logging features provide auditors with detailed visibility into all data access activities.
Comments	It is not free.

2. Group: Specific Repository Systems

8. Name of service / tool	**Integrated Data Repository Toolkit (IDRT)**
Contact person	idrt@imise.uni-leipzig.de
Webpage	http://idrt.imise.uni-leipzig.de/IDRT-II/
Country	Germany
Cross-country	Yes
Short description (only a few sentences)	The Open Source software i2b2 provides a translational research platform for storing biomedical data and querying these data with a user-friendly interface

	it is lacking user-friendly tools for installation and configuration, the import of source data and the creation of a comprehensive navigational structure (i2b2 ontology) To close these gaps, the Integrated Data Repository Toolkit (IDRT), consisting of three software tools, has been created. The i2b2 Wizard provides a shell GUI for the installation and configuration of i2b2 instances, projects and users.
Type of activity (project, service, collaboration, platform, etc.)	The Integrated Data Repository Toolkit (IDRT) was conceived to address both the issue of setup and administration as well as the ability to import clinical data in several standard formats, including terminologies.
Modules/components included	**i2b2 Wizard:** Automates many core aspects of setting up and administrating an i2b2 instance. **IDRT-Import-Tool:** a Java GUI with an i2b2 server browser and convenient configuration wizards for starting the ETL. Therefor an extensive set of configurable Talend Open Studio ETL-jobs has been developed to import fact data from common standard formats such as CSV, CDISC ODM and SQL databases into the i2b2 system. The ETL process supports the automatic derivation of i2b2 ontologies from metadata (such as column names in CSV files) and the automatic replacement of flat code lists with standard terminological hierarchies (e.g. ICD-10) in the i2b2 ontology. We've also implemented import jobs for German standard terminologies and the German "§21" benchmarking data set. **i2b2-Ontology-Editor (IOE):** A Java GUI to edit i2b2 Ontologies.
Data included	Medical data
Research use cases (including collaborations)	**Clinical Data Warehouse:** • implemented at Erlangen University Hospital, providing i2b2 as an additional query, frontend for research applications • import hospital benchmarking datasets into i2b2 • data sets ranging from 42,000 to 47,000 patients with around 53,000 cases and up to 1,3 million individual observations **Research Data Repository**

	• The trial database of the Competence Network for Congenital Heart Defects – available in CDISC ODM files – was integrated with i2b2 • The source data inherent ODM Ontology hierarchy is used for composing selections and filters on the data • Over 16000 data sets with roughly 80 concepts were imported and partially combined with imaging metadata **Translational Research Unit** • Developed for the Clinical Research Group 241 • focusses on the longitudinal course of psychosis • trial data from the local team in Göttingen can be accessed as CDISC ODM from the secuTrial System (iAS GmbH) • The ODM file contains data for over 700 patients and defines about over 1500 items, resulting in over 1.8 million patient facts • use i2b2 as an easy tool for quality assurance and data management
moComments	

9. Name of service/tool	**Local EGA - European Genome-phenome Archive**
Contact person	Pascal Kahlem, Scientific Network Management, S.L., Sant Just Desvern (Barcelona), Spain, pkahlem@gmail.com
Webpage	https://github.com/elixir-europe/human-data-local-ega https://www.elixir-europe.org/use-cases/human-data
Country	Central EGA database in UK
Cross-country	European
Short description (only a few sentences)	EGA provides a service for the permanent archiving and distribution of personally identifiable genetic and phenotypic data resulting from biomedical research projects. Data at EGA was collected from individuals whose consent agreements authorise data release only for specific research use to bona fide researchers. Strict protocols govern how information is managed, stored and distributed by the EGA project. So-called 'local EGAs' will store national sensitive data locally and integrate their metadata globally in the central EGA. While the local storage will conform to the national requirements for personal data protection, the metadata integration in EGA will ensure the national data are discoverable at international level.
Type of activity (project, service, collaboration, platform, etc.)	Storing, archiving and sharing of personally identifiable genetic and phenotypic human data resulting from biomedical research projects.
Modules/components included	The Local EGA project is divided into several microservices: Db = Postgres database mq = RabbitMQ message broker with appropriate accounts, exchanges, queues and bindings Inbox SFTP server is acting as a dropbox, where user credentials are in central EGA keyserver = handles the encryption/decryption keys workers connect to the keyserver (via SSL) and do the actual re-encryption task ? vault moves files from the staging area to the vault storage, including a verification step afterwards Local EGA Docker is installed. The EGA Database interaction server provides most database functionality for the EGAPRO PostgreSQL database. The project is created using Netbeans using ant. Ant target "package-for-store" creates the packaged Jar file containing all libraries. There are no further dependencies; everything is packaged in

	the Jar file. Servers use Netty as framework. Client REST calls use Resty.
Data included	Human genetic and phenotypic data
Research use cases (including collaborations)	ELIXIR has a Human Data Use Case for local EGA. The Use Case takes the European Genome-phenome Archive (EGA) as its primary data source, access to which is controlled. The EGA allows an authorised user to search sequenced material, patient samples stored in biobanks, and the metadata around patients (their illnesses, treatments, outcomes). It also queries national search engines on behalf of the users. Datasets can then be downloaded into an EGA compatible cloud or cluster local to the researcher. The Human Data Use Case provides a framework for the secure submission, archiving, dissemination and analysis of human biomedical data across Europe by local EGAs. The Use Case develops for this purpose a portable submission toolkit (Local-EGA) to deposit sensitive human data locally (and comply with national guidelines for storing that data). It enables data reuse across national boundaries. One must be part of an ELIXIR Node, to set up a local instance of the EGA with metadata from the main EGA. This will allow people to search both the local and the central EGA at once. A Submission REST API is developed to submit data to a local EGA programmatically. Local EGAs can store metadata from the central EGA, which allows to use the local EGA to search both the main and local EGA. One can also search and retrieve information from the local EGA by using the local-EGI API, so one can build own services based on the data available. This framework ensures that services handling human data comply with the General Data Protection Regulation (GDPR).
Comments	

10. Name of service / tool	**De-identified Clinical Data Repository (DCDR)**
Contact person	University of Washington Institute of Translational Health Sciences 850 Republican St Box 358051 Seattle, WA 98109 Phone: (206) 221-1234 FAX: (206) 616-9250
Webpage	https://www.iths.org/investigators/services/bmi/dcdr/
Country	
Cross-country	
Short description (only a few sentences)	The DCDR is a de-identified data repository that contains a subset of data from various UW Medicine clinical systems. Please see the data dictionary page for details about the data elements contained in the DCDR.
Type of activity (project, service, collaboration, platform, etc.)	The tool provides a graphical query interface that allows the definition of criteria and returns an aggregate count or a summary of the patients who meet the criteria. An example query in this interface would be of the form: “Provide me a count of patients 18-34 with a diagnosis of diabetes mellitus, who were discharged live within the past six months.” **Accessing the DCDR** 1. Request DCDR access by filling out the Access Request Form. We will notify you within two business days of your access status. 2. Once you receive your access approval, go to the DCDR Web Interface to access the tool. 3. Go through self-directed training modules to learn how to query the DCDR.
Modules/components included	The DCDR is a cohort identification/feasibility estimation tool. The interface to the DCDR is a secure web-based query tool (currently powered by i2b2) that allows researchers to individually query the data.
Data included	
Research use cases (including collaborations)	
Comments	

11. Name of service / tool	**Integrating Data for Analysis, Anonymization and SHaring (iDASH)**
Contact person	Email: idash@ucsd.edu(link sends e-mail) Phone: 858-246-1794
Webpage	https://idash.ucsd.edu/
Country	USA
Cross-country	Yes
Short description (only a few sentences)	Integrating Data for Analysis, Anonymization and SHaring (iDASH) is one of the National Centers for Biomedical Computing (NCBC) under the NIH Roadmap for Bioinformatics and Computational Biology. Founded in 2010, the iDASH center is hosted on the campus of the University of California, San Diego and addresses fundamental challenges to research progress and enables global collaborations anywhere and anytime. Driving biological projects motivate, inform, and support tool development in iDASH. iDASH collaborates with other NCBCs and disseminates tools via annual workshops, presentations at major conferences, and scientific publications.
Type of activity (project, service, collaboration, platform, etc.)	iDASH is supported by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant U54HL108460. **Participating Institutions** •UCSD Jacobs School of Engineering •San Diego Supercomputer Center (SDSC)(link is external) •California Institute for Telecommunications and Information Technology (Calit2)(link is external) •UCSD School of Medicine •UCSD Skaggs School of Pharmacy and Pharmaceutical Sciences •UCSD Medical Center •UCSD Moores Cancer Center **What is MIDAS?** MIDAS is the open-source platform from Kitware upon which the iDASH repository is based. Customizations have been made to make this instance of MIDAS that are specific for iDASH purposes.
Modules/components included	**iDASH offers the following resources within a secure cyberinfrastructure:** •Secure, privacy-preserving data repository •Open-source software •Schemas, models, and algorithms

Data included	Medical Data
Research use cases (including collaborations)	**Genomic Data Privacy Protection Using Compressive Sensing, University of Oklahoma -Tulsa** **iDASH Data Repositories** An important part of the information infrastructure provided by iDASH is to provide a single, comprehensive set of facilities to explore, navigate, analyze, and combine different forms of information provided by different data sources, within the bounds of privacy restrictions. iDASH is designed to be scalable and extensible so that developers can integrate the heterogeneous data from the national biomedical, clinical, and informatics communities. Developed as an open, community-serving, crowd-sourcing resource, the iDASH team is collaborating with biomedical, behavioral, and quantitative researchers to establish the nation’s most robust data repository for high-quality collections of data. This rich repository of medical data includes images and text accompanied by meta-data. **Data users** Without a MIDAS account, you can download any data contained in a public folder. Private folders require a MIDAS account and approval from the community owner for access. **How do I upload data into the iDASH repository?** To upload data into the iDASH repository, you first need to obtain a MIDAS account. To obtain a MIDAS account, follow the appropiate DUA procedures or contact idash@ucsd.edu(link sends e-mail) to help you get started.
Comments	

Group 3: Complementary tools and solutions

Generic Tools & Solutions

12. Name of service / tool	**Microsoft Azure**
Contact person	Germany 0800-627-1035
Webpage	https://azure.microsoft.com/en-us/
Country	USA
Cross-country	Yes
Short description (only a few sentences)	Microsoft Azure (formerly Windows Azure) is a cloud computing service created by Microsoft for building, testing, deploying, and

	managing applications and services through a global network of Microsoft-managed data centres.
Type of activity (project, service, collaboration, platform, etc.)	It provides software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS) and supports many different programming languages, tools and frameworks, including both Microsoft-specific and third-party software and systems.
Modules/components included	Operating system: Linux, Microsoft Windows Azure is productive for developers: Get your apps to market faster. Azure integrated tools, from mobile DevOps to serverless computing support your productivity. Build the way you want to, using the tools and open source technologies you already know. Azure supports a range of operating systems, programming languages, frameworks, dAzure is the only consistent hybrid cloud Azure is the only consistent hybrid cloud: Build and deploy wherever you want with Azure, the only consistent hybrid cloud on the market. Connect data and apps in the cloud and on-premises—for maximum portability and value from your existing investments. Azure offers hybrid consistency in application development, management and security, identity management, and across the data platform. Azure is the cloud for building intelligent apps: Use Azure to create data-driven, intelligent apps. From image recognition to bot services, take advantage of Azure data services and artificial intelligence to create new experiences—that scale—and support deep learning, HPC simulations, and real-time analytics on any shape and size of data. Azure is the cloud you can trust: Ninety percent of Fortune 500 companies trust the Microsoft Cloud. Join them. Take advantage of Microsoft security, privacy, transparency, and the most compliance coverage of any cloud provider.
Data included	Storage Services provides REST and SDK APIs for storing and accessing data on the cloud.
Research use cases (including collaborations)	There many companies (for example: Adobe, HP, Maersk, Siemens and Coca cola) that uses some Azure products.
Comments	

13. Name of service / tool	**Amazon Web Services (AWS) & Cloud Computing**
Contact person	
Webpage	https://aws.amazon.com/
Country	U.S.
Cross-country	Yes
Short description (only a few sentences)	AWS has the services to help you build sophisticated applications with increased flexibility, scalability and reliability
Type of activity (project, service, collaboration, platform, etc.)	Financial Services, Digital Marketing, Media and Entertainment, Gaming, Enterprise Applications, **Healthcare & Life Sciences**, Government, Nonprofit, Education, Automotive, Manufacturing, Power & Utilities
Modules/components included	Some developers and management tools
Data included	Any type of data
Research use cases (including collaborations)	**Healthcare & Life Sciences** **Genomics**: Bring your tool sets to free data sets in the AWS Cloud, and accelerate time to results. **Biotech & Pharma:** Easily deploy globally for any type of development, lab, or manufacturing workload. **Healthcare Providers & Insurers:** Accelerate your digital transformation and enable advanced healthcare analytics.
Comments	

14. Name of service / tool	**Orion Metadata Harvester**
Contact address / person, if available	Contact person: Suzanne Clark +15105813086 Company: OrionIC inc 4165 Amyx Ct. Hayward, CA 94542 United States +1 510 4848706
Webpage of the tool	https://www.oriongovernance.com/products/metadata-harvester/ http://www-304.ibm.com/partnerworld/gsd/solutiondetails.do?&solution=54531&lc=en
Country it is used	USA
Cross-country use?	**Yes.** Countries/regions available for distribution: Americas, United States, Asia Pacific, Australia, Europe, United Kingdom
Short description of the tool	Orion Governance has extensive, global, experience providing clients with data governance and data lineage solutions, with a long and successful history of untangling the structured metadata world, and now the addition of unstructured assets (Excel, SharePoint, Pdf, etc.).
Type of activity the tool supports (project, service, collaboration, platform, etc.)	**Metadata Sources:** Structured Data(SQL, ETL, COBOL/JCL), Big Data(Hadoop, HiveDB); Unstructured Data(Email, SharePoint, FileNet)
Modules / architecture / components included	The Orion Metadata Harvester is deployed as a software appliance. The appliance is a Linux based machine which can run on a VMWare ESX environment or in a bare metal windows based environment running VMWorkstation. The solution is modular, scalable and extendable.
What data is stored in the tool	The platform includes scanners to extract **metadata**, auto schedulers, data lineage, impact analysis, mapping recommendation engine and integration to just about any data source.
Research use cases / projects / studies the tool is used (including collaborations)	**Solution areas:** • Financial Services/Banking, • **Healthcare and Pharmaceutical (Physician Office Automation, Clinical information systems, Health information systems),** • Cross industry (Surveillance and security, Governance, Risk and Compliance)
Comments	Orion Metadata Harvester is not free. The following links is protocol that is used by it. This protocol name is Open Archives Initiative Protocol for Metadata Harvesting (**OAI-PMH**) https://en.wikipedia.org/wiki/Protocol_for_Metadata_Harvesting **OAI-PMH uses:** • Some commercial search engines use OAI-PMH to acquire more resources. Google initially included support for OAI-

	PMH when launching sitemaps, however decided to support only the standard XML Sitemaps format in May 2008. • In 2004, Yahoo! acquired content from OAIster (University of Michigan) that was obtained through metadata harvesting with OAI-PMH. • Wikimedia uses an OAI-PMH repository to provide feeds of Wikipedia and related site updates for search engines and other bulk analysis/republishing endeavors. Especially when dealing with thousands of files being harvested every day, OAI-PMH can help in reducing the network traffic and other resource usage by doing incremental harvesting. • NASA's Mercury: Metadata Search System uses OAI-PMH to index thousands of metadata records from Global Change Master Directory (GCMD) every day. • The mod_oai project is using OAI-PMH to expose content to web crawlers that is accessible from Apache Web servers

15. Name of service / tool	**DataTags**
Contact address / person, if available	Hardvard University email: privacytools-info@seas.harvard.edu
Webpage of the tool	https://datatags.org/ https://privacytools.seas.harvard.edu/datatags
Country it is used	
Cross-country use?	
Short description of the tool	DataTags, a suite of tools to help researchers share and use sensitive data in a standardized and responsible way. Proper handling of human subject's data requires knowledge of relevant federal and state data privacy laws, applicable data sharing agreements, best practices for confidentiality and security, and available mechanisms for privacy protection. The goal of DataTags is to help researchers who are not legal or technical experts navigate these considerations and make informed decisions when collecting, storing, and sharing privacy-sensitive data.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	This is an ongoing project that is in collaboration with the IQSS Dataverse team.
Modules / architecture / components included	**Data-tagging Tools** help a data holder choose an appropriate DataTag and sharing policy for a dataset are currently available and under development. **Six standardized DataTags levels:** Blue, Green, Yellow, Orange, Red, Crimson
What data is stored in the tool	DataTags
Research use cases / projects / studies the tool is used (including collaborations)	**HIPAA Research Archive** The Health Information Portability and Accountability Act (HIPAA) is a 1996 federal statute that authorized the U.S. Department of Health and Human Services to establish privacy rules governing individually identifiable health information, rules that specify with whom and how physicians, hospitals, and insurers may share a patient's medical information. **Multinational Corporation Archive** Multinational corporations acquire data from many diverse sources, under a multitude of agreements, and are subject to many laws, regulations, and business practices.

	Global Research Repository Global Research Repository supports a full range of sensitive data from any researcher in the world.
Comments	

16. Name of service / tool	**Secure folder system**
Contact address / person, if available	Indiana University
Webpage of the tool	https://kb.iu.edu/d/bbox, https://uits.iu.edu/box
Country it is used	USA
Cross-country use?	
Short description of the tool	The IU Box service provides a simple, secure way to share and store files and folders online. Box consolidates your content in a single location, easily accessible from anywhere, on any device. You can create files and folders, share them using a direct link, invite others to collaborate, and continue to revise and review your content. Though similar in appearance to other consumer services such as Dropbox, Box can directly integrate with existing IU systems (e.g., accounts, CAS for single sign-on with your IU username and passphrase), security, and contractual protections. Your Box account quota is unlimited.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	• View files of many types in your browser or mobile device, including images and audio/video. • Access content through all major current web browsers (i.e., Internet Explorer, Firefox, Chrome, Safari) and through mobile devices running iOS, Android, and BlackBerry • Access through Microsoft Office applications (Windows only)
Modules / architecture / components included	• Share files and folders while controlling the level of access others have, with a range of permissions from view-only to full editing and collaboration rights • Comment on files • Create simple workflows using assigned tasks • Sync files between your desktop and other devices, and access them even when offline
What data is stored in the tool	Any type of data
Research use cases / projects / studies the tool is used (including collaborations)	Used in Indiana University
Comments	**Box at Indiana University is not appropriate for storing or sharing most types of institutional data classified as Critical.** However, with

	certain additional security measures you may be able to use IU Box with some data that contain protected health information (PHI) regulated by the Health Insurance Portability and Accountability Act of 1996 (HIPAA).

17. Name of service / tool	**Docker**
Contact address / person, if available	Docker, Inc. 144 Townsend St. San Francisco, CA 94107 (415) 941-0376 Get help with Docker support@docker.com
Webpage of the tool	https://docs.docker.com/
Country it is used	
Cross-country use?	Yes
Short description of the tool	Docker provides a way to run applications securely isolated in a container, packaged with all its dependencies and libraries.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	Docker containers are lightweight by design and ideal for enabling **microservices application development**.
Modules / architecture / components included	The first step with Docker is **to modernize the existing application** portfolio. **Packaging existing apps** into containers immediately improves security, reduce costs, and gain cloud portability. Cloud migration, multi-cloud or hybrid cloud infrastructure requires frictionless **portability of applications**. Docker packages applications and their dependencies together into an isolated container making them portable to any infrastructure. Eliminate the “works on my machine” problem once and for all.
What data is stored in the tool	Secret data
Research use cases / projects / studies the tool is used (including collaborations)	**1. A TMF project** initiated in late 2015 has explored the evaluation and deployment of Docker technology for **biomedical research**. A broad variety of tools can be migrated to Docker containers, and are available to researchers and IT decision-makers via a Docker Hub. These “dockable” resources include i2b2, tranSMART, the TMF PID generator, plus tools for the MOSAIC project at **Greifswald University Hospital**, and OpenClinica, an open-source software solution. **2. Bioinformatics software.** A segment of the bioinformatics industry is leveraging Docker container images to build BioShadock, a custom Docker registry for bioinformatics tools and software. As community members explain in this paper, the goal of BioShadock is to provide a concentrated repository of bioinformatics programs without relying on generic Docker registries. In addition, because the software is available as containers, it can be installed easily.

	3. Simplified Deployment of **Health Informatics Applications** by Providing Docker Images. **4. Distributed Applications and Microservices** You can use containers to create distributed applications by breaking apart your application into independent tasks or processes (e.g., microservices). **5. Batch/ETL Jobs** You can use containers for batch and ETL jobs by packaging the job into a container and deploying it into a shared cluster. **6. Continuous Integration and Continuous Deployment** You can use containers for continuous integration and deployment because Docker provides a system for image versioning. Use of application containers and workflows for **genomic data analysis** **7. NASA’s Land Information System (LIS).** The LIS traditionally has been difficult to install due to complex software dependencies. With Docker, LIS installation has become much easier.
Comments	

18. Name of service / tool	**ELIXIR Beacons**
Contact address / person, if available	Contact: serena.scollen@elixir-europe.org
Webpage of the tool	https://www.elixir-europe.org/about-us/implementation-studies/beacons https://github.com/ga4gh-beacon/beacon-elixir
Country it is used	European Counties
Cross-country use?	Yes
Short description of the tool	ELIXIR Beacons provide discovery services for genomic data in the European Genome-phenome Archive (EGA) and in ELIXIR Nodes, using the Beacon technology developed by the Global Alliance for Genomics and Health (GA4GH).
Type of activity the tool supports (project, service, collaboration, platform, etc.)	The **GA4GH Beacon** is an open sharing platform that allows any genomic data centre in the world to make its data discoverable. Users can ask Beacons straightforward questions like, 'Do any of these data resources have genomes with this allele at that position?' The search result informs a researcher as to whether making a data access request is required for their research, saving valuable time and resource.
Modules / architecture / components included	Docker image available at: https://hub.docker.com/r/egacrg/beacon/ It includes the Elixir Beacon application, already deployed and running, and a PostgreSQL database with some sample data. The JAR file is located at /tmp folder with the default configuration (for further information see section Elixir Beacon, the main project). The database used is called elixir_beacon_dev and the default user and password are microaccounts_dev and r783qjkldDsiu.
What data is stored in the tool	Human DNA, RNA data
Research use cases / projects / studies the tool is used (including collaborations)	The first stage of the project (2015-2016) focused on establishing Beacons within ELIXIR Nodes and resulted in six ELIXIR Beacons in:

	• ELIXIR Belgium • ELIXIR Finland (the first ELIXIR Beacon connected to the ELIXIR AAI to demonstrate registered data access level) • ELIXIR France • ELIXIR Switzerland: Beacon Array Map, DIPG Beacon • ELIXIR Sweden • European Genome-phenome Archive (EGA) In January 2017, ELIXIR extended its collaboration with GA4GH to further drive the development and implementation of the Beacon technology across ELIXIR Nodes. The main goals for 2017 are: • Establish the network of ELIXIR Beacons • Develop new features • Add security measures to attract stakeholders with more sensitive data sets while minimising risks to individual privacy, • Increase strategic partnering with national data owners to enable data flow to the Beacon service
Comments	

19. Name of service / tool	**Dutch Techcenter for Life Sciences**
Contact address / person, if available	**Mail address:** PO Box 8500 3503 RM Utrecht The Netherlands T: +31 (0)85 – 30 30 711
Webpage of the tool	https://www.dtls.nl/
Country it is used	Netherlands
Cross-country use?	
Short description of the tool	The Dutch Techcentre for Life Sciences (DTL) is a public-private partnership of more than 50 life science organisations in the Netherlands. The majority of Dutch universities and university medical centres are DTL partners and a growing number of companies are joining the organisation. DTL is organised as a network of experts and policymakers affiliated with the DTL partners. This DTL network is supported by a small facilitating team located in Utrecht. (Read more about the organisation.) DTL connects scientists, data experts, technical experts, and trainers that are specialised in a variety of high-end wet lab and data technologies, and working in life science domains ranging from health to nutrition, agro, biotech, and biodiversity. Together, these professionals interconnect and improve their research infrastructure to enable cost-effective cross-technology life science research in national and international collaboration.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	**Data** DTL actively promotes FAIR Data Stewardship of life science information, within its partnership and in close collaboration with its international partners. FAIR stands for 'Findable, Accessible, Interoperable, and Reusable'. The FAIR Data principles act as an international guideline for high quality data stewardship.
Modules / architecture / components included	**ELIXIR-NL** DTL acts as the Dutch node of ELIXIR, the European data infrastructure for the life sciences. ELIXIR unites Europe's leading life science organisations in managing and safeguarding the increasing volume of data generated by publicly funded research. ELIXIR-NL has three focus areas: compute and storage infrastructure, data interoperability, and training & education. Central to many of ELIXIR-NL's activities are the FAIR Data principles (i.e., research data should be Findable, Accessible, Interoperable, and Reusable for both humans and computers).
What data is stored in the tool	DTL connects scientists, data experts, technical experts, and trainers that are specialised in a variety of high-end wet lab and data technologies, and working in **life science domains ranging from health to nutrition, agro, biotech, and biodiversity**.

Research use cases / projects / studies the tool is used (including collaborations)	**Initiatives** DTL has been set up by its partners to interconnect local infrastructures. To this end, DTL and its partners are actively involved in several large-scale research infrastructure initiatives. • About large-scale research infrastructures • Health-RI • Netherlands X-omics initiative • News about large-scale research infrastructures
Comments	

20. Name of service / tool	**The European Data Portal**
Contact address / person, if available	Call us between 09:30 - 17:30 (CET) EN: +352 31 44 01-448 FR: +352 31 44 01-449 email: help@europeandataportal.eu
Webpage of the tool	https://www.europeandataportal.eu/
Country it is used	
Cross-country use?	Yes
Short description of the tool	The European Data Portal harvests the metadata of Public Sector Information available on public data portals across European countries. Information regarding the provision of data and the benefits of re-using data is also included.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	Datasets, Catalogues, Metadata Quality, Licensing Assistant, SPARQL Manager, Statistics
Modules / architecture / components included	Factsheets and reports
What data is stored in the tool	Datasets, catalogues, statistics
Research use cases / projects / studies the tool is used (including collaborations)	https://www.europeandataportal.eu/en/using-data/use-cases
Comments	

21. Name of service / tool	**Open Metadata Registry**
Contact person	Diane Hillmann, Cornell University, dih1@cornell.edu
Webpage	http://metadataregistry.org/
Country	USA
Cross-country	
Short description (only a few sentences)	The Open Metadata Registry is a fundamental piece of technical infrastructure for the Semantic Web. The Registry is available openly and to all who wish to use its services. It provides a means to identify, declare and publish through registration metadata schemas (element/property sets), schemes (controlled vocabularies) and Application Profiles (APs). In addition to supporting registration of schemes, schemas and APs for consumption and use by human and machine agents, the Open Registry supports the machine mapping of relationships among terms and concepts in those schemes (semantic mappings) and schemas (crosswalks). Thus, the Registry will support the key goals of metadata discovery, reuse, standardization and interoperability locally and globally.
Type of activity (project, service, collaboration, platform, etc.)	It was originally built to support the National Science Digital Library (NSDL),
Modules/components included	The Registry used as its inspiration the open-source Dublin Core Metadata Initiative (DCMI) Registry. The Registry extended the original DCMI goals to support: (1) the automated creation and maintenance of schemas and application profiles; and (2) the submission of schemas and schemes to a registry workflow for review and publication. Server was moved from Rackspace to Digital Ocean and from one server to several smaller ones, using Ubuntu 14LTS, nginx (openresty actually), and PHP5.5.
Data included	Property Vocabularies are registered. A set of properties or classes (also called here an Element Set or an Element Vocabulary) starts with a description of the set or vocabulary as a collective whole. A new vocabulary form can be invoked by clicking on the (Add) link and clicking on the Element Sets link will produce a link of already registered Element Sets.
Research use cases (including collaborations)	Example: used by NASA Mars Science Laboratory for Navigation Camera file
Comments	

22. Name of service / tool	**Aristotle metadata registry**
Contact person	Diane Hillmann, Cornell University, dih1@cornell.edu
Webpage	http://aristotle-metadata-registry.readthedocs.io/en/master/ https://github.com/aristotle-mdr/aristotle-metadata-registry https://registry.aristotlemetadata.com/ (open instance of registry)
Country	Australia
Cross-country	
Short description (only a few sentences)	Aristotle-MDR is an open-source metadata registry as laid out by the requirements of the ISO/IEC 11179 specification. It represents a new way to manage and federate content built on and extending the principles of leading metadata registries. The code of Aristotle is completely open-source, building on the Django web framework and the 11179 standard, allowing agencies to easily run their own metadata registries while also having the ability to extend the information model and tap into the permissions and roles of ISO 11179.
Type of activity (project, service, collaboration, platform, etc.)	The core of the Aristotle Metadata Registry is designed to conform to the models described within ISO/IEC 11179-3, However this mono-repo includes a number of standards-based extensions that provide additional functionality or new metadata types. built upon the Django web framework and the mature model of the ISO/IEC 11179 standard. As such agencies can easily run their own metadata registries while also having the ability to extend the information model and tap into the permissions and roles of ISO/IEC 11179. The Aristotle Open Metadata Registry is an implementation of the Aristotle Metadata Registry, an open-source registry framework, available for all users. Browse the Aristotle-MDR GitHub page for instructions on downloading and customising it.
Modules/components included	Aristotle-MDR is free open-source software and contributions are welcome on front-end web development, back-end server development, translation and content creation. SQLite database, PythonAnywhere, linux, run aristotle_mdr.install.easy for installation. To test Aristotle, there is an included \`Dockerfile\`. This will Use the \`/aristotle_mdr/example_mdr/\` django settings file and install Aristotle-MDR and all requirements, create an SQLite Database and Whoosh search index inside the Container and collect the necessary static files.
Data included	
Research use cases (including collaborations)	Aristotle-MDR is used to show usability of open government data. For this metadata from Aristotle Metadata Registry is connected with data from data.gov.au. Evolve SBR - Australian Government Standard Business Reporting Data Dictionary. Prosper Canada Financial

	Literacy Indicator Registry - Registry of peer-reviewed indicators tracking financial literacy outcomes for Prosper Canada, OCASI Indicator Registry - Metadata registry for the Ontario Council of Agencies Serving Immigrants
Comments	

23. Name of service / tool	**Open Metadata Repository Services (OMRS)**
Contact person	Mandy Chessell, IBM, mandy_chessell@uk.ibm.com
Webpage	https://cwiki.apache.org/confluence/pages/viewpage.action?pageId=70258803
Country	UK
Cross-country	international
Short description (only a few sentences)	Traditional metadata management technology tends to centralize metadata into a single repository. The OMRS enables the integration of metadata that is distributed amongst a number of metadata repositories either: through a call interface which is provided by an OMRS connector, using notifications that broadcast changes to metadata in a repository that other repositories can subscribe to and via linked data URLs that enable a metadata entity to have a relationship with a metadata entity in a different repository.
Type of activity (project, service, collaboration, platform, etc.)	The Open Metadata Repository Services (OMRS) enable metadata repositories to exchange metadata.
Modules/components included	To join an open metadata repository cohort, a metadata repository must support the following OMRS integration methods. •Support for an OMRS repository connector to allow open metadata API calls to the repository to create, query, update and delete metadata stored in the repository. The OMRS connectors support the Open Connector Framework (OCF) to provide a call interface to the metadata repositories. •Support for the OMRS event notifications that are used to synchronize selective metadata between the metadata repositories. •Support for OSLC linked data relationships to allow relationships between metadata entities that happen to reside in different metadata repositories. OMRS Operational Services Supports the administration services for the Open Metadata Repository Services (OMRS). OMRS Configuration Factory Generates default values for the Open Metadata Repository Services (OMRS) configuration. OMRS Audit Log Reads and writes audit log messages. OMRS Archive Manager Reads and loads the content from open metadata archives. Governance Client provides Java classes that call the Open Metadata Access Services (OMAS) REST APIs (requires the IP address and port number of the deployed OMASs) and OMAS Message Helpers to build payloads for the OMASs's Topics. Open Metadata and Governance (OMAG) Server provides a server runtime for Open Metadata and Governance Components that can be selectively configured to support different integration patterns for tools using the Caller and Adapter Integration Patterns. Open Metadata and Governance Discovery Server provides a server runtime for discovery services that analyze data resources to augment the metadata about them, or report exceptions to quality or protection standards. These discovery services plug into the Open Discovery

	Framework (ODF). The Discovery Toolkit provides the interfaces for building ODF compliant discovery services.
Data included	Metadata
Research use cases (including collaborations)	Used by CIBC, SAS, Microsoft, Oracle, Informatica, ING., Waterline, RBC
Comments	

24. Name of service / tool	**Neo4j**
Contact person	1-855-636-4532 US info@neo4j.com +44 808 189 0493 UK uk@neo4j.com Please email: Germany vertrieb@neo4j.com +33 (0) 8 05 08 03 44 France ventes@neo4j.com
Webpage	https://neo4j.com/product/ and https://neo4j.com/
Country	
Cross-country	Yes
Short description (only a few sentences)	**A Graph Platform Reveals and Persists Connections** The graph platform takes a connections-first approach to data. It broadens a company's ability recognize the importance of persisting relationships and connections through every transition of existence: from idea, to design in a logical model, to implementation in a physical model, to operation using a query language and to persistence within a scalable, reliable database system. The foundation of representing connected data is known as a graph. Neo4j's Graph Platform is built around the Neo4j native graph database: • The Neo4j native graph database supports transactional applications and graph analytics • Graph analytics help data scientists gain new perspectives on data • Data integration expedites distilling tabular data and big data into graphs • The Cypher graph query language is the bridge to big data analytic tooling • Graph visualization and discovery help communicate graph technology benefits throughout the organization • Enterprise architecture underlies and supports massive graph data
Type of activity (project, service, collaboration, platform, etc.)	
Modules/components included	
Data included	

Research use cases (including collaborations)	**General Use Cases:** Fraud Detection, Identity and Access, Knowledge Graph, Master Data Management, Network and IT Operations; Privacy, Risk and Compliance; Recommendation Engine, Social Network **Neo4j Use Cases in Life Sciences and Healthcare:** In biology, biochemistry, pharmaceuticals, healthcare and other life sciences, you know that you work with highly-connected information. Unfortunately, many scientists still use relational databases and spreadsheets as their daily tools. Here we want to present you with an alternative. Managing, storing and querying connected information is natural to a graph database like Neo4j. Learn how your research and practitioner colleagues utilized Neo4j to draw new insights or just be more efficient in their daily work. It started a while time ago in 2012 with a workshop at the University of Ghent bringing together people from the field with graph database experts. Now we want to take it to the next level by providing you with a platform to present your projects and paper both here and on our blog, and giving you the opportunity to connect with other Neo4j users in your field. **An Integrated Data Driven Approach to Drug Repositioning Using Gene-Disease Associations** An alternative database approach for management of SNOMED CT and improved patient data queries **Knowledge.Bio:** A Web application for exploring, building and sharing webs of biomedical relationships mined from PubMed Representing and querying disease networks using graph databases **Graph Databases in Life and Health Sciences; we had topics covering:** Neo4j in metaproteonomics Graph databases in cancer research Project collaboration networks and recommendations Detailed studies of citation graphs Connecting protein databases in a large graph model “Reactome” database of human protein interaction pathways **Life Sciences and Healthcare Accelerator Program** The Neo4j Life Sciences and Healthcare Accelerator Program is designed to help researchers and practitioners in life sciences and healthcare-related sciences make sense of their data using Neo4j. Whether you are analyzing genome data, combining protein databases, investigating drug interactions or supporting practitioners with research

	or clinical information processing we want to help you find insights in connected (meta-)data. **Using Graph Technology to Fight Cancer** https://neo4j.com/news/using-graph-technology-to-fight-cancer/ Graph databases excel in tasks related to search and recommendation because they not only store information about individual things, but also the relationships between those things. This capability allows users to ask questions that were previously not possible with traditional database technologies. The data relationships stored in the graph database can express the nature of each connection (e.g. drug family, type of cancer targeted) and capture any number of qualitative or quantitative facts about that relationship (e.g. optimal dosage level, treatment success rate, effectiveness against mutations, and date brought to market). Once loaded into a graph database, an entirely new set of relationship-based questions can be asked, opening up new possibilities. **Data Management in Systems Biology and Medicine:** https://neo4j.com/blog/data-management-systems-biology-medicine/ Integrating Data for Translational Medicine Informatics: Incorporating a Graph Database into the eTRIKS Data Architecture The Data Model: Protein Framework Metabolic-Centric Framework: Converting XML Data Exploring Conditional Probabilities Using Neo4j for Data Visualization
Comments	

25. Name of service / tool	**Comprehensive Knowledge Archive Network (CKAN)**
Contact person	
Webpage	https://ckan.org/
Country	
Cross-country	Yes
Short description (only a few sentences)	CKAN is a powerful data management system that makes data accessible – by providing tools to streamline publishing, sharing, finding and using data.
Type of activity (project, service, collaboration, platform, etc.)	
Modules/components included	DATASTORE: The CKAN DataStore extension provides an ad hoc database for storage of structured data from CKAN resources. METADATA: A CKAN portal provides a rich set of metadata for each dataset. SEARCH AND DISCOVERY: CKAN provides a rich search experience which allows for quick 'Google-style' keyword search as well as faceting by tags and browsing between related datasets. VISUALIZATION: CKAN's data previewing tool has a host of powerful features for previewing data stored in the DataStore. PUBLISH AND MANAGE DATA: An intuitive web interface allows publishers and curators to easily register, update and refine datasets.
Data included	Any type of data
Research use cases (including collaborations)	**The Australian Federal Open Data Portal:** Launched in 2013, that provides access to public datasets from government. Initially, the open data portal was based on WordPress. Eventually, to provide a better experience both to data publishers and end users, it was decided to migrate to a specialised data publishing application. **Link Digital** was chosen to make this possible via the open source **CKAN platform**. **A CKAN + Drupal open data portal by Civity:** Dataplatform.nl is a CKAN + Drupal open data portal. Providing a central repository, the data portal allows each city to have its own website, e.g. utrecht.dataplatform.nl or denhaag.dataplatform.nl.
Comments	Open source

26. Name of service / tool	**Dataverse**
Contact person	Email: support@dataverse.org
Webpage	https://dataverse.org/
Country	
Cross-country	Yes
Short description (only a few sentences)	Dataverse is an open source web application to share, preserve, cite, explore, and analyze research data. It facilitates making data available to others, and allows you to replicate others' work more easily. Researchers, journals, data authors, publishers, data distributors, and affiliated institutions all receive academic credit and web visibility.
Type of activity (project, service, collaboration, platform, etc.)	Funded by Harvard with additional support from the Alfred P. Sloan Foundation, National Science Foundation, National Institutes of Health, Helmsley Charitable Trust, IQSS's Henry A. Murray Research Archive, and many others.
Modules/components included	**Academic Credit:** By depositing data into Dataverse, which can be customized or embedded into a website with our Theme + Widgets feature, researchers make their datasets more discoverable to the scientific community. Widgets are available at the Dataverse and dataset level and can be embedded in any website to help others find a scholar's datasets more easily. **Data Citation:** Dataverse standardizes the citation of datasets to make it easier for researchers to publish their data and get credit as well as recognition for their work. When you create a dataset in Dataverse, the citation is generated and presented automatically. **Data Management:** By depositing research data in a Dataverse repository (including Harvard Dataverse) researchers can fulfill funding agency requirements for data management plans. Creation of a data management plan is a "best practice" for research projects that involve the collection or dissemination of data.
Data included	Each dataverse contains datasets, and each dataset contains descriptive metadata and data files (including documentation and code that accompany the data). Dataverse stores the raw data content extracted from such files in plain text, TAB-delimited files. The metadata information that describes this content is stored separately, in a relational database, so that it can be accessed efficiently by the application. For the purposes of archival preservation it can be exported, in plain text XML files, using a standardized, open DDI Codebook format.
Research use cases (including collaborations)	The Institute for Quantitative Social Science (IQSS) collaborates with the Harvard University Library and Harvard University Information Technology organization to make the installation of the Harvard

	Dataverse openly available to researchers and data collectors worldwide from all disciplines, to deposit data. All research data files in the Harvard Dataverse repository are stored in an Amazon S3 bucket. All content placed in that bucket is immediately replicated to a second S3 bucket in a different, isolated availability zone. After seven days in this second bucket, all files are moved into Glacier, Amazon's cloud data archiving service for long-term backup storage.
Comments	

Specific Tools & Solutions

□□□□Name of service / tool	**MOLGENIS**
Contact address / person, if available	Contact: Dr. Morris Swertz Dept. of Genetics, CB50 University Medical Center Groningen P.O. Box 30001 9700 RB GRONINGEN The Netherlands email: m.a.swertz@rug.nl
Webpage of the tool	https://molgenis.github.io/ http://www.molgenis.org/wiki/WikiStart
Country it is used	
Cross-country use?	
Short description of the tool	MOLGENIS is a modular web application for scientific data. MOLGENIS was born from molecular genetics research (and was called 'molecular genetics information system') but has grown, thanks too many sponsors and contributors, to be used in many scientifc areas such as biobanking, rare disease research, patient registries and even energy research. MOLGENIS provides researchers with user friendly and scalable software infrastructures to capture, exchange, and exploit the large amounts of data that is being produced by scientific organisations all around the world. To get an idea of what the software can do, visit our MOLGENIS YouTube channel.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	MOLGENIS is modular, having all kinds of extension modules to store and interact with your data.
Modules / architecture / components included	MOLGENIS is two things: 1. For biologists MOLGENIS is a suite of web databases for genotype, phenotype, QTL and analysis pipelines. 2. For bioinformaticians MOLGENIS is software generator to rapidly build web databases that make your biologists happy.

What data is stored in the tool	biobanking, rare disease research, patient registries and even energy research data.
Research use cases / projects / studies the tool is used (including collaborations)	**MOLGENIS applications:** • xQTL Workbench for multi-level QTL mapping (project, publication) • Dystrophic Epidermolysis Bullosa (deb-central) mutation database (project, publication) • eXtensible Genotype and Phenotype database (XGAP) (project, publication) • Design of Genetical Genomics Experiments (designGG)] (project, publication) • BBMRI-NL biobank catalogue (project) • MAGE-TAB microarray gene experiment object model (MAGETAB-OM)] (project, demo) • Pheno-OM Phenotype observation model (project, demo) • Mouse Resource Browser (MRB) project] (project, publication) • MOLGENIS as data wrapper in Taverna (publication) • Animal observation database (AnimalDB) (project) • Nordic GWAS control database (project, publication) • GWAS Central curation tool (project) • Finnish disease database (FINDIS) (project) • Bacterial microarrays database (MOLGEN-IS) (publication) • Human Metabolic Pathway Database (project, publication)
Comments	

□□□□Name of service / tool	**TranSMART**
Contact address / person, if available	i2b2 tranSMART Foundation 401 Edgewater Place, Suite 600 Wakefield MA 01880 USA contact@i2b2transmart.org
Webpage of the tool	http://transmartfoundation.org/ https://github.com/transmart https://wiki.transmartfoundation.org/display/transmartwiki/Getting+Support
Country it is used	USA
Cross-country use?	Yes
Short description of the tool	The tranSMART knowledge management and high-content analysis platform is a flexible software framework featuring novel research capabilities. It enables analysis of integrated data for the purposes of hypothesis generation, hypothesis validation, and cohort discovery in translational research. tranSMART bridges the prolific world of basic science and clinical practice data at the point of care by merging multiple types of data from disparate sources into a common environment. The application supports data harmonization and integration with analytical pipelines.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	**The tranSMART platform makes it possible for scientists to:**

	• Develop, test and refine research hypotheses • Search multiple data sets for potential drug targets, pathways and biomarkers • Compare data from proteomics, metabolomics and other “omics” studies • Contrast patterns of gene expression in healthy and diseased individuals and tissue samples • Investigate correlations between genotype and phenotype in clinical trial data • Mine pre-clinical data for insights into the biology of human disease • Study genetic and environmental factors involved in human disease • Display data visually using a graphical interface • Stratify clinical data into molecular subtypes of a specific disease • Collaborate across academic and corporate research sectors
Modules / architecture / components included	Longitudinal data Support Cross Study Support Support for High Volume Variant Data Upgrade path/i2b2 integration Continuation of SmartR or other plugin visualization/analytical tools (e.e. Spotfire) Support for standards and internal proprietary ontologies
What data is stored in the tool	Genetic and phenotypic data, and assessing their analytical results in the context of published literature and other work

Research use cases / projects / studies the tool is used (including collaborations)	• Patient Stratification and association analysis • Cohort comparison • Data Curation at Pfizer • Accelerated Cure Project for Multiple Sclerosis • Collaboration, analysis, customization in preclinical oncology • Predictive toxicology • Getting the most from public data • Multi-omics data analysis (Bio-IT World)
Comments	

29. Name of service / tool	**TraIT**
Contact address / person, if available	service desk at: servicedesk@ctmm-trait.nl CTMM-TraIT service desk by telephone on +31 (0)88 1167500
Webpage of the tool	http://www.ctmm-trait.nl/ http://www.ctmm.nl/en/projecten/translational-research-it-trait/translationele-research-it-trait?set_language=en
Country it is used	translational research projects in the Netherlands
Cross-country use?	TraIT already supports over 475 research projects spanning almost 4000 individual researchers.
Short description of the tool	TraIT enables integration and querying of information across the four major domains of translational research: clinical, imaging, biobanking and experimental (any-omics) with a particular focus on the needs of multi-center projects.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	Data Integration and Analysis Deployment and User Support
Modules / architecture / components included	Clinical Trails
What data is stored in the tool	**Clinical Research Data:** Clinical Imaging, Digital Pathology, Biosample Data, Experimental DataDomains
Research use cases / projects / studies the tool is used (including collaborations)	• The TraIT project started as an initiative from the Center for Translational Molecular Medicine (CTMM). • CTMM and TI Pharma have merged and continue as Lygature as of January 1st 2016. • At the end of 2016, the TraIT Foundation (Stichting TraIT) was formed. • Since its inception in October 2011 the TraIT project grew from 11 partners to 32 participating organisations including all Dutch University Medical Centers, charities such as the Dutch Cancer and Heart Foundations, and a variety of private partners.
Comments	

30. Name of service / tool	**TSD (Tjenester for Sensitive Data)**
Contact address / person, if available	tsd-contact@usit.uio.no
Webpage of the tool	http://www.uio.no/english/services/it/research/sensitive-data/
Country it is used	Norway
Cross-country use?	
Short description of the tool	The TSD - Service for Sensitive Data, is a platform to collect, store analyze and share sensitive data in compliance with the Norwegian regulation regarding individuals privacy. TSD is used by researchers working at UiO(University of Oslo) and in other public research institutions (UH-sector, university hospitals etc.). The TSD is primarily an IT-platform for research even if in some case it is used for clinical research and commercial research.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	To import and export files in/out TSD one must use the TSD file exchange tool, sometimes called File Lock. Sensitive data must be encrypted before import/export. In special cases it is also possible to import data directly from disk. One of the following file transfer tools should be installed before one can start importing/exporting files to TSD: sftp, FileZilla or winSCP. The TSD servers have a number of software packages installed by default and optional packages are available together with the possibility to make additional requests for software.
Modules / architecture / components included	**Windows servers:** Windows 2012 servers with SAS, Matlab, stat, R, and more. **Linux servers:** Redhat 6.0 with Libre (open) Office and R. **Storage:** All projects are issued with a basic amount of storage space which can be expanded if needed. **High performance computing:** Projects can apply for access to the HPC cluster Colossus. **Infrastructure:** The solution is run on dedicated computers in a separate location in USITs machine room where only USITs operational personnel have access. To achieve complete separation of project environments running on the same hardware, we use RHEV KVM as a hypervisor. This means that a physical computer can be divided into several separate virtual computers which for all intents and purposes are working independently. **Security** •All access from external networks demands two-step authentication. •The computers are hardened more than normal. •All user management is done per environment. This means that the security does not depend on the users regular UiO account.

	•All changes in access rights is done with a written approval from the project administrator (in TSD 2.0 this can often be done in minID). •Dedicated storage, encrypted backups and encrypted communication is used. •Encryption keys are generated with a unique set of keys for each project/environment. These are stored on paper in a safe in two separate locations. Data transfer in and out of the system is done via a special purpose file staging service.
What data is stored in the tool	All type of human sensitive data
Research use cases / projects / studies the tool is used (including collaborations)	System is use in University of Oslo in Norway.
Comments	**For installing additional Python packages please have a look at:** http://www.uio.no/english/services/it/research/hpc/abel/help/software/Python%202.html **You will find the prices for all standard TSD services here:** http://www.uio.no/english/services/it/research/sensitive-data/access/prices/index.html

31. Name of service/tool	**TRYGGVE**
Contact person	project manager: Antti Pursula, antti.pursula@csc.fi
Webpage	https://wiki.neic.no/wiki/Tryggve
Country	Nordic countries
Cross-country	Yes (Denmark, Finland, Norway, and Sweden)
Short description (only a few sentences)	The Tryggve project is an EGA and ELIXIR collaboration with the aims to provide researchers with a trusted set of services for sharing and analysing sensitive medical data, across Nordic countries. TRYGGVE establishes a Nordic platform for the collaboration with sensitive data. It is a cross-country activity with NeIC and ELIXIR nodes as partners. Its approach is to utilize and connect existing capacities and services in the Nordic countries, using 3 Mio Euro, for 3 years and 100 PMs. The use case covers different types of solutions, pools different data from different countries, provides secure transfer of data, and moves the computation and the software into containers
Type of activity (project, service, collaboration, platform, etc.)	Tryggve is a three-year project to establish a Nordic platform for collaboration on sensitive data, funded by NeIC and the ELIXIR nodes in Denmark, Finland, Norway, and Sweden.
Modules/components included	Components of the TRYGGVE framework are: • Computerome at DTU, DK • ePouta secure cloud at CSC, FI • Mosler service, NBIS, SE (though, Mosler is rather small) • TSD2 at USIT, NO • Everything is connected with ELIXIR node
Data included	Genetic data Sensitive medical data
Research use cases (including collaborations)	Research use cases cover the TransNordic gene analysis environment, for the pooling of data and the Scandinavian Genetic Collaboration. In a Schizophrenia research use cases, the Mosler remote desktop is connected to ePouta for data analysis and provision of a data platform. The EDA use case consists of EGA/local EGA, collects associated metadata, using DAC for access, and REMS provided by ELIXIR FI. The central EGA is used for search and authorization and brings data from EGA into ePouta's secure environment for analysis. In another use case, researcher at SurfSara access ePouta and TSC (HPC). Here the legally compliant transfer of data is the issue. The data of clinical cohorts are a mixture of consented data with other data, e.g.,

	diabetes study data from hospitals in Finland and Germany, etc., which can be used for data mining.
Comments	**Tryggve offers following use cases** • Accessing and moving research data • Support in meeting legal and ethical requirements • Secure data analysis platforms • Share data with your colleagues • Software installations • Access to data archives

32. Name of service / tool	**UK Data Service Secure Lab**
Contact address / person, if available	University of Essex Wivenhoe Park Colchester Essex CO4 3SQ +44 (0)1206 872143
Webpage of the tool	https://www.ukdataservice.ac.uk/use-data/secure-lab
Country it is used	England
Cross-country use?	Yes
Short description of the tool	The UK Data Service Secure Lab provides secure access to data that are too detailed, sensitive or confidential to be made available under the standard End User Licence or Special Licence. Data accessed in this way cannot be downloaded. Once researchers and their projects are approved, they can analyse the data remotely from their organisational desktop, or by using our Safe Room. We provide access to statistical and office software to make remote analysis and collaboration secure and convenient.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	We also provide Microsoft Office and LaTeX software to allow researchers to write up their results within the Secure Lab environment.
Modules / architecture / components included	Access is via a web-based interface that uses secure encrypted Citrix Virtual Private Network technology. The data are never downloaded. We provide users of the Secure Lab with a familiar Windows environment and the statistical tools that they require to achieve their analyses. We currently provide the following statistical software in the Secure Lab: R, Stata, SPSS 19 (with linear regression package), MLwiN, Mplus, ArcGIS, GeoDa
What data is stored in the tool	It provides researchers with access to sensitive and confidential **business, social and economic microdata**. Its growing collection of datasets which are derived from survey, administrative and transaction sources include: • Productivity data from the Annual Respondents Database • Innovation data from the UK Innovation Survey • Geospatial data from the Labour Force Survey, Understanding Society • Sensitive data about childhood development **Data Depositors for Secure Access Data**

	- Office for National Statistics - University of Essex, Institute for Social and Economic Research - Centre for Longitudinal Studies - Department for Education - Department for Communities and Local Government - Scottish Centre for Social Research - Department for Transport - Department for Work and Pensions Department of Energy and Climate Change
Research use cases / projects / studies the tool is used (including collaborations)	It is a data service. Any registered user can download, order, or analyse data online.
Comments	

33. Name of service / tool	**DataSHIELD**
Contact address / person, if available	Professor Paul Burton, University of Bristol Office OF20 Oakfield House, Oakfield Grove, Clifton BS8 2 BN +44 (0) 117 3310072 p.burton@bristol.ac.uk
Webpage of the tool	https://www.datashield.ac.uk/
Country it is used	England
Cross-country use?	Yes. Canada, China, Six European countries.
Short description of the tool	DataSHIELD is an infrastructure and series of R packages that enables the remote and non-disclosive analysis of sensitive research data. Users are not required to have prior knowledge of R.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	**Multi-site DataSHIELD** This infrastructure is appropriate for the co-analysis of harmonised individual level data held at multiple locations. Each data location installs the server-side DataSHIELD infrastructure that holds a snapshot of the harmonised data to be co-analysed. One of the locations also installs and manages the DataSHIELD client portal, the mechanism by which users are authenticated to send analysis commands within the DataSHIELD infrastructure. **Single-site DataSHIELD** This infrastructure is used to enable analysis of individual level data held at one location. In this case, the data server-side DataSHIELD infrastructure is installed in addition to the DataSHIELD client portal.
Modules / architecture / components included	Client connects using virtual machine. Client system can be Linux, Windows and Mac. The virtual servers require 1.5GB RAM each and about 1GB hard-disk space each.
What data is stored in the tool	Data is stored in Opal Datawarehouse / Datawarehouses.
Research use cases / projects / studies the tool is used (including collaborations)	Current DataSHIELD Pilots: **InterConnect** (MRC Epidemiology Unit, Cambridge) is developing a global collaborative network for **diabetes** and **obesity** research. The **BioSHaRE-EU Healthy Obese Project** for the federated analysis of ten European studies including data from the National Child Development Study. The **BioSHaRE-EU Environmental Core Project** for the federated analysis of data from six European studies including UK Biobank.

	C SPIRIT (University of Sherebrooke) intra-uterine determinants of child health and development and on perinatal health services in Quebec, Canada (3 studies) and Shanghai, China (1 study). **ENPADASI** (German Institute of Human Nutrition) will pilot DataSHIELD to deliver an open access research infrastructure that will contain data from a wide variety of nutritional studies, ranging from mechanistic/interventions to epidemiological studies including a multitude of phenotypic outcomes that will facilitate combined analyses in the future.onsortia Setting up DataSHIELD Pilots.
Comments	

34. Name of service / tool	**EHR4CR (Electronic Health Records for Clinical Research)**
Contact address / person, if available	Project coordinator Dr. Mats Sundgren Principal Scientist, Global Clinical Development AstraZeneca, Sweden mats.sundgren@astrazeneca.com
Webpage of the tool	http://www.ehr4cr.eu/
Country it is used	**France**, **Germany**, **Poland**, **Switzerland** and **United Kingdom**.
Cross-country use?	Yes
Short description of the tool	The EHR4CR platform is an open IT platform that unlocks the information stored in Electronic Health Records for improving clinical research while fully respecting patient privacy and ensuring a high level of security.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	The EHR4CR project (2011-2016) with a budget of +16 million Euros, has involved 35 academic and private partners (10 pharmaceutical companies) and is one of the largest of the IMI PPPs in this area. It was part-sponsored by the European Commission through the Innovative Medicines Initiative (IMI).
Modules / architecture / components included	The EHR4CR platform specifies a Common Information Model (CIM). Additionally, tool builders can consider all platform information to be structured according to this CIM. EHR data is not natively structured according to this CIM. The combination of an ETL (Extract, Transform and Load) step and the EHR4CR semantic integration layer takes care of the transformation between the local information models and the EHR4CR CIM.
What data is stored in the tool	Query work on Electronic health records.
Research use cases / projects / studies the tool is used (including collaborations)	The consortium included 11 hospital sites in **France**, **Germany**, **Poland**, **Switzerland** and the **United Kingdom**.
Comments	

35. Name of service / tool	**Aircloak**
Contact address / person, if available	Contact us at solutions@aircloak.com Aircloak Berlin Gormannstrasse 14 10119 Berlin Germany
Webpage of the tool	https://www.aircloak.com/
Country it is used	Germany
Cross-country use?	Yes
Short description of the tool	Aircloak allows for immediate, safe and legal sharing or monetisation of sensitive datasets. Aircloak sits between rich databases and outside analysts – like a filter for personal data – enabling much higher fidelity analytics than any existing anonymisation or data masking solution.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	Works with your existing SQL and NoSQL databases Aircloak Insights is installed in front of and connects to your existing database servers.
Modules / architecture / components included	**Anonymising Filter** Aircloak's first-in-class solution is installed between the existing "primary-use" database and untrusted "secondary use" analysts and applications. It filters queries and answers to ensure user anonymity using Aircloak Insights **Aircloak Insights** is a transparent proxy between analysts and the sensitive data they want to work with. Analysts query the system with SQL, like they are used to, and Aircloak Insights writes a query tailored to the backend storing the data.
What data is stored in the tool	Sensitive clinical data
Research use cases / projects / studies the tool is used (including collaborations)	**Medical Research:** Sharing medical data for research is a costly and risky process. Huge opportunities for improved medical research are lost because medical institutions leave data locked-up in protected silos. Aircloak enables: • Keep full control over sensitive and valuable data and only ever grant anonymising access to results. • Improved quality of analytics by preventing the loss of data accuracy normally associated with traditional forms of anonymization. • Compliance with medical research anonymization requirements, thus eliminating the need for new patient opt-ins. **Banking:** Aircloak Insights' approach to instant compliant analytics is instrumental in allowing banks to comply with complex regulations.

	Automotive: Modern car manufacturers are data companies. Sensitive and personal information is continuously collected through wide arrays of smart sensors. **Service providers:** Service providers amass user data through every single interaction their customers perform. Everything from making a call, browsing the web, or streaming an on-demand show leaves a digital trace available to the service provider.
Comments	

36. Name of service / tool	**The Secure Data Vault (SDV)**
Contact address / person, if available	Contact Jim at jim.morris@datasecurityadvisorygroup.org
Webpage of the tool	https://ubiquity.acm.org/article.cfm?id=3081882
Country it is used	
Cross-country use?	
Short description of the tool	The Secure Data Vault (SDV) is an approach to protecting the most sensitive data from malware and insider exploits. The vision of an SDV as a microservices architecture built on a minimal trusted computing base (TCB) that would include the seL4 microkernel configured as a hypervisor with the formally verifiable properties of being both malware-secure and insider-secure.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	Services at its own databases
Modules / architecture / components included	Every microservice (that needs one) contains its own database. Each microservice is an independent, simple process virtual machine (for example, a Java Virtual Machine, or JVM). The seL4 microkernel, configured as a hypervisor, manages the multiple JVMs, including all inter-process communication between the microservices. An SDV doesn't need a platform VM (virtual machine) or an OS-level VM. A process VM (e.g. a JVM) has almost everything an SDV microservice needs, and the system services implemented on top of the seL4 microkernel will supply what the JVM needs but doesn't have. Each microservice is a virtual machine because the microservices need to be encapsulated and isolated from each other for better security.
What data is stored in the tool	All type of data can be stored.
Research use cases / projects / studies the tool is used (including collaborations)	Protecting the most sensitive data from malware and insider exploits. **We could not find medical usage**.
Comments	The following **10 rules** are guidelines for building, deploying, and using an SDV. **Rule Number 1:** An SDV must be structured as a microservices architecture that can be decomposed into microservices that are simple enough that they can be individually verified using formal methods. **Rule Number 2:** Each microservice must have only one purpose (or goal or function). This aids in the development of relatively simple microservices.

	Rule Number 3: An SDV must guarantee (by formal verification, if possible) that it is immune to external infection by malware. **Rule Number 4:** User authentication (e.g., user "log-ins") must rely on multi-factor biometrics. **Rule Number 5:** The difficulty and inconvenience of retrieving sensitive data from an SDV should be directly proportional to the amount of sensitive data to potentially be retrieved. **Rule Number 6:** The number of humans (at least two in the case of least-sensitive data) required to simultaneously submit a request for retrieval for any and all sensitive data from an SDV should be directly proportional to the sensitivity level of the data. This is similar to the Two-Man Rule, which requires two humans to take simultaneous action in order to launch a nuclear weapon. Inconvenient, but necessary. **Rule Number 7:** Security (access controls) must be automated as much as possible, and it should be extremely difficult for humans to disable or relax security protections. **Rule Number 8:** Manual security (access-control) configuration required by humans should be kept to a minimum, or, if possible, should even be non-existent. **Rule Number 9:** Software updates should update a microservice in its entirety, the update must be secure, and the update should be accomplished remotely "over the air," as is done with all smartphones today. **Rule Number 10:** Network communication (over a path that contains insecure computers) between microservices that exist on two physically different computers in the network must be encrypted.

37. Name of service / tool	**Jisc DataVault**
Contact address / person, if available	email: info@datavaultplatform.org
Webpage of the tool	http://libraryblogs.is.ed.ac.uk/jiscdatavault/ The full Project Plan can be read online: https://docs.google.com/document/d/1k2XHlNBGR7sM6XBfyICIeGguoJc5uP3JJwJhrtgRvhI/edit?usp=sharing Documentation and issue tracking into the DataVault GitHub repository https://github.com/DataVault/datavault
Country it is used	England
Cross-country use?	
Short description of the tool	The Jisc Research Data Spring the universities of Manchester and Edinburgh have been continuing to develop DataVault. Both institutions are currently planning local implementations, whilst working together to continue develops the software.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	A joint University of Edinburgh and University of Manchester Jisc Data Spring project.
Modules / architecture / components included	Tasks on the DataVault development list: – BagIt Libraries, resolve issues with BagIt libraries: removing empty directories, renaming files, creating manifests in memory – Stand alone packager, to package deposits external to the DataVault web application – Verification of deposits, e.g. checking that the number of files and filesizes are correct – Deposits via API, for browser upload, requires further investigation into authentication and chunking of files – User Roles/Groups/Sharing Archives, definitions of roles and implementation within the DataVault – Closing Vaults, in what circumstances are vaults closed? – Integrations with Pure and Dropbox
What data is stored in the tool	sensitive data
Research use cases / projects / studies the tool is used (including collaborations)	The universities of Manchester and Edinburgh have been continuing to develop DataVault. Both institutions are currently planning local implementations.

Comments	

38. Name of service / tool	**European Genome-phenome Archive**
Contact address / person, if available	Help Desk: ega-helpdesk@ebi.ac.uk
Webpage of the tool	https://www.ebi.ac.uk/ega/home https://github.com/elixir-europe/human-data-local-ega
Country it is used	
Cross-country use?	Yes
Short description of the tool	The European Genome-phenome Archive (EGA) is designed to be a repository for all types of sequence and genotype experiments, including case-control, population, and family studies. We will include SNP and CNV genotypes from array based methods and genotyping done with re-sequencing methods.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	The European Genome-phenome Archive (EGA) is available at the European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG). The EGA provides a service for the permanent archiving and distribution of personally identifiable genetic and phenotypic data resulting from biomedical research projects. Data at EGA was collected from individuals whose consent agreements authorise data release only for specific research use to bona fide researchers. Strict protocols govern how information is managed, stored and distributed by the EGA project. For controlled access data, the EGA will provide the necessary security required to control access, and maintain patient confidentiality, while providing access to those researchers and clinicians authorised to view the data. In all cases, data access decisions will be made by the appropriate data access-granting organisation (DAO) and not by the EGA. The DAO will normally be the same organisation that approved and monitored the initial study protocol or a designate of this approving organisation.
Modules / architecture / components included	It is a combined system of module, architecture and components.
What data is stored in the tool	**What are Studies:** Studies are experimental investigations of a particular phenomenon or trait. **What are Datasets?:** Datasets are defined file collections, whose access is governed by a Data Access Committee (DAC).
Research use cases / projects / studies the tool is used (including collaborations)	The EGA will serve as a permanent archive that will archive several levels of data including the raw data (which could, for example, be re-analysed in the future by other algorithms) as well as the genotype calls provided by the submitters.

Comments	

39. Name of service / tool	**ELIXIR Use Case for human data research**
Contact address / person, if available	Contact Pascal Kahlem pkahlem@gmail.com
Webpage of the tool	https://www.elixir-europe.org/use-cases/human-data https://www.rd-alliance.org/system/files/documents/elixir-human-data-rdap10.pdf
Country it is used	European Countries
Cross-country use?	Yes
Short description of the tool	The Use Case takes the European Genome-phenome Archive (EGA) as its primary data source, access to which is controlled. The EGA allows an authorised user to search sequenced material, patient samples stored in biobanks, and the metadata around patients (their illnesses, treatments, outcomes). It also queries national search engines on behalf of the users. Datasets can then be downloaded into an EGA compatible cloud or cluster local to the researcher. The Human Data Use Case extends and generalises the system of access authorisation and secure data transfer developed in the EGA. It aims to provide a framework for the secure submission, archiving, dissemination and analysis of human biomedical data across Europe.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	**Provides standardised tools to discover and access human data** **Local-EGAs for metadata sharing:** By extending the use of Local-EGAs the Use Case is increasing the amount of human data that is discoverable. Local-EGAs store metadata from the main EGA, which will allow you to use the local EGA to search both the main and local EGA. You can also search and retrieve information from the Local-EGA by using the Local-EGI API, so you can build your own services based on the data available. In addition, the main EGA will gather the metadata from all the data submitted to Local-EGAs, so a search at the main EGA will allow you to find data located across all Local-EGAs. **Beacon project:** The Human Data Use Case is working with the Global Alliance for Genomics and Health (GA4GH) to use the beacon discovery service for resources across ELIXIR. The Beacon service provides a simple way to make data discoverable. You can query the lightweight metadata provided by a data resource (a 'beacon') to ask questions like 'Does this dataset have genomes with this allele at that position?' and get a 'Yes' or 'No' answer. **Regulating access to sensitive data:** the Use Case is working with the Compute Platform to use the ELIXIR Authentication and Authorization Infrastructure (AAI) for ELIXIR resources. The AAI is a system that allows you to have a single identity across a range of different services, so you can use the same log-in for each service. The ELIXIR AAI also contains a Resource Entitlement Management System (REMS). This provides a way that you can request access online to a restricted data resource, and a Data Access Committee (DAC) for that resource can

	review your application. If you are granted access, you can the log in to the resource using the AAI (which verifies your right to access the data).
Modules / architecture / components included	**Provides a sustainable infrastructure for storing, coordinating and distributing human data** The infrastructure is based on the European Genome-phenome Archive (EGA), tranSMART and Galaxy. Researchers will use the EGA to store their raw data, tranSMART to collate different data sets for preliminary analysis, and a Galaxy cloud service for further analysis. **Local-EGA:** The Use Case is also developing a portable submission toolkit (Local-EGA). This will allow you to deposit sensitive human data locally (and comply with national guidelines for storing that data) but enable data reuse across national boundaries. If you are part of an ELIXIR Node, you can set up a local instance of the EGA with metadata from the main EGA. This will allow people to search both your local and the main EGA at once. **Submission REST API:** the Use Case is developing an API that you can use to submit data to a Local-EGA programmatically.
What data is stored in the tool	European Genome-phenome Archive (EGA) as its primary data source
Research use cases / projects / studies the tool is used (including collaborations)	**It is a use case.**
Comments	

40. Name of service / tool	**Computerome**
Contact address / person, if available	Computerome services, please contact: Ali Syed, E-mail: alisyed@dtu.dk or phone: (+45) 60 90 46 46. For support: hpc@bio.dtu.dk
Webpage of the tool	http://www.computerome.dtu.dk/
Country it is used	Denmark
Cross-country use?	Yes
Short description of the tool	The Danish National Life Science Supercomputing Center, Computerome is a HPC Facility specialized for Life Science. Users include Research groups from all Danish Universities and large international research consortiums as well as users from industry and the public Health Care Sector. They all benefit from the fast, flexible and secure infrastructure and the ability to combine different types of sensitive data and perform analysis. Computerome is physically installed at the DTU Risø campus and managed by a strong team of specialists from DTU. Computerome is the official supercomputer of ELIXIR Denmark, a member of ELIXIR, the European infratructure for biological information.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	**Services**

	• Separate storage with HIPAA-compliant auditing • Private copies of reference databases like TCGA and Ensemble • Batch Queuing system – to ensure SLA • Wide selection of preconfigured and regularly updated scientific tools • Support for both Windows and Linux • Extreme Scalability (from 1 CPU VM to 1000+ CPU) • Complete separation from the internet or monitored and filtered one way traffic • Access to specialised services e.g. GPU nodes, off-site redundant backup • Data Analysis & visualization services • Dynamic Pipelines • Cloud bursting – cloud in a box. • Data collection to analysis service (BGI online) • CLC Bio workbench • Cloud based GATK • Analysis of human gene expression and regulation • Large genomic pipelines in a portable and reproducible manner • Data collection • Hadoop & Spark • R studio suite
Modules / architecture / components included	**Secure Cloud** Using cloud technology as a delivery mechanism for HPC you get a lot of advantages and you hide the underlying hardware/Software complexities. **With Cloud architecture on top of HPC you can:**

	• Apply data management and automation • Set up private secure virtual supercomputers • Enable sophisticated security and governance setup • Adapt to local statutory acts • Scale resources hour for hour **Get your own virtuel supercomputer with secure cloud** With our secure cloud solution you can get your own private secure virtual supercomputer at Computerome. Your data will be secured in a private cloud, accesable from anywhere, but only to whoom you choose. This has the advantage that both private companies and scientists, who can’t afford to buy, update and maintain their own supercomputer, are now able to obtain all the benefits of a supercomputer. **Security measures with secure cloud:** • Smart phones or programmable hardware tokens • Import/export is under strict control • No open connection to the internet • All administration happens from inside • Separation between projects • Hardened gateway and firewall • Encrypted backup, one key per project • Sys-admins are single users (traceability) **COMPUTEROME CLOUD TODAY:** • Accessible from anywhere through proper mechanisms • Secure private clusters • Virtual clusters • Cloud bursting capabilities • Security, isolation and access control in compliance with statutory acts
What data is stored in the tool	different types of sensitive data

Research use cases / projects / studies the tool is used (including collaborations)	**Use Cases** Computerome has already processed over two million tasks. In this link, 7 scientists explain how the supercomputer has helped them with their research. http://www.computerome.dtu.dk/use-cases
Comments	

41. Name of service / tool	**CREDENTIAL**
Contact address / person, if available	E-Government Innovation Center EGIZ IAIK, TU-Graz, Inffeldgasse 16a, A-8010 Graz Austria
Webpage of the tool	https://www.egiz.gv.at/en/research/25-CREDENTIAL-A-Framework-for-Privacy-Preserving-Cloud-Based-Data-Sharing
Country it is used	
Cross-country use?	
Short description of the tool	A Framework for Privacy-Preserving Cloud-Based Data Sharing.Data sharing -- and in particular sharing of identity information - plays a vital role in many online systems. While in closed and trusted systems security and privacy can be managed more easily, secure and privacy-preserving data sharing as well as identity management becomes difficult when the data are moved to publicly available and semi-trusted systems such as public clouds. CREDENTIAL is therefore aiming on the development of a secure and privacy-preserving data sharing and identity management platform which gives stronger security guarantees than existing solutions on the market. The results will be showcased close to market-readiness through pilots from the domains of eHealth, eBusiness, and eGovernment, where security and privacy are crucial. From a technical perspective, the privacy and authenticity guarantees are obtained from sophisticated cryptographic primitives such as proxy re-encryption and redactable signatures.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	
Modules / architecture / components included	
What data is stored in the tool	
Research use cases / projects / studies the tool is used (including collaborations)	
Comments	

42. Name of service / tool	**RD Connect**
Contact address / person, if available	Project Manager: Libby Wood libby.wood@newcastle.ac.uk T: +44 191 241 8621 RD-Connect coordination team Institute of Genetic Medicine Newcastle University International Centre for Life Newcastle upon Tyne NE1 3BZ United Kingdom
Webpage of the tool	http://rd-connect.eu/
Country it is used	
Cross-country use?	Yes
Short description of the tool	RD-Connect is an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. To help researchers study rare diseases, RD-Connect links different data types - omics (e.g. genomics), clinical information, patient registries and biobanks - into a common resource. RD-Connect enables scientists and clinicians around the world to analyse genomics data and share them with other researchers. By making data accessible beyond the usual institutional and national boundaries, RD-Connect speed up research, diagnosis and therapy development to improve the lives of patients with rare diseases.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	The RD-Connect project received six-years funding (2012 - 2018) from the European Union, under the Seventh Framework Programme (FP7) grant, to create a unique infrastructure for rare disease research free to use by scientists and clinicians in Europe and around the world. RD-Connect is a collaborative work of partners based in Europe and beyond.
Modules / architecture / components included	• Genome-phenome analysis platform • Bioinformatic tools • Data linkage • Ethical, Legal and Social Issues (ELSI) • Patient engagement
What data is stored in the tool	Omics data, Phenotypic data, Biosamples data

Research use cases / projects / studies the tool is used (including collaborations)	Analyse genomics data and share them with other researchers
Comments	

43. Name of service / tool	**The European Network for Cancer Research in Children and Adolescents (ENCCA)**
Contact address / person, if available	Contact us E-mail: office@siope.eu Tel: +32 2 775 02 12
Webpage of the tool	https://www.siope.eu/encca/
Country it is used	
Cross-country use?	Yes
Short description of the tool	The European Network for Cancer Research in Children and Adolescents (ENCCA) is a network of excellence comprising 34 high-level European research institutes and organisations, recognised for their distinction in paediatric oncology and dedication to improving the treatment of children and adolescents suffering from cancer.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	SIOPE is in charge of the communication and dissemination activities of the project, as well as of the management and facilitation activities. Funded by the FP7 (7th Framework Programme for research of the European Commission) for 5 years, the project will last until the end of 2015. ENCCA efficiently integrates all stakeholders (including parent and patient advocates) by enhancing their collaboration and providing them with common tools and approaches to solve the current bottlenecks in testing new therapeutic strategies for rare cancers in a vulnerable age group. In an effort to facilitate access to therapies and standards of care across Europe, the ultimate aim of ENCCA is to create a sustainable "European Virtual Institute" for clinical and translational research in childhood and adolescent cancers.
Modules / architecture / components included	
What data is stored in the tool	Cancer data
Research use cases / projects / studies the tool is used (including collaborations)	
Comments	

44. Name of service / tool	**EMIF (European Medical Information Framework)**
Contact address / person, if available	CONTACT: Project coordinator Bart Vannieuwenhuyse Janssen Pharmaceutica NV BVANNIEU@its.jnj.com
Webpage of the tool	http://www.imi.europa.eu/projects-results/project-factsheets/emif
Country it is used	
Cross-country use?	Yes
Short description of the tool	The EMIF project aims to develop a common information framework of patient-level data that will link up and facilitate access to diverse medical and research data sources, opening up new avenues of research for scientists. To provide a focus and guidance for the development of the framework, the project will focus initially on questions relating to obesity and Alzheimer's disease.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	EMIF-Platform Project. EMIF will build an integrated, efficient Information Framework for consistent re-use and exploitation of available patient-level data to support novel research. This will support data discovery, data evaluation and then (re)use. EMIF-Platform will develop an IT platform allowing access to multiple, diverse data sources. The EMIF-Platform will make the data available for browsing and allow exploitation in multiple ways by the end user. EMIF-Platform will initially be able to, on its own, leverage data on around 40 Million European adults and children by means of federation of healthcare databases and cohorts from 6 different countries (DK, IT, NL, UK, ES, EE), designed to be representative of the different types of existing data sources (population-based registries, hospital-based databases, cohorts, national regis To guide the development of the framework, the team will initially focus on two key research issues: 1. identifying the mechanisms that make some people more susceptible to dementias (such as Alzheimer's disease) than others; 2. determining which individuals with obesity are most likely to develop complications such as diabetes. Obesity and dementia are two of the greatest healthcare challenges of our time; EMIF's work will pave the way for new diagnostic tools and treatments to help patients with these conditions. Looking to the future, additional research areas may be added to the framework through future IMI Calls for proposals.tries, biobanks, etc.). http://www.emif.eu/about/emif-platform
Modules / architecture /	EMIF Deliverable 5.1: Establishing functional data warehousing and management http://www.emif.eu/assets/e/m/emif_d5_1_data_warehousing_management_exec_summary_website.pdf

components included	
What data is stored in the tool	patient-level data initially on questions relating to obesity and Alzheimer’s disease
Research use cases / projects / studies the tool is used (including collaborations)	
Comments	

45. Name of service / tool	**Open Source Registry System for Rare Diseases in the EU (OSSE)**
Contact person	Marita Muscholl, University Medical Center, Mainz, Germany, muscholl@uni-mainz.de
Webpage	http://osse-register.de
Country	Germany
Cross-country	European countries
Short description (only a few sentences)	Open-source software for the creation of patient registers. OSSE primarily provides a registry toolbox that allows for the definition of forms for longitudinal and medical core data and of the corresponding data schema by means of a registry editor. Common data elements, which can be chosen to build the forms, are defined within a metadata repository (MDR) following ISO/IEC 11179. Interoperability between different OSSE registries is achieved by a distributed search infrastructure taking into account data ownership and privacy aspects. The central search broker allows specified search queries in all OSSE-compliant registries based on the existing MDR items.
Type of activity (project, service, collaboration, platform, etc.)	Development of a minimum data set and documents with regard to compliance with data protection regulations. Recruitment of reference registers and subsequent conversion of an existing register to a new IT system with the help of OSSE and networking of existing registers with the help of the so-called OSSE bridgehead while retaining the previously used register software. The integration of a central metadata directory facilitates integration of data coming from different OSSE registries because the specifications of all Rare Disease related records used in the registers are viewable through the MDR.
Modules/components included	The OSSE Registry provides role-based access control, plausibility checking for the validity of entered data, versioning of collected data, and initial workflow support using various statuses of forms and role-dependent allowed state transitions. The data import and export can be configured via an integrated graphical user interface. For improved visibility, each register should register in a register of registers (RoR). The principle of "distributed search" requires that patient data should not leave the local registry, even if the patient's consent is given, as there are clear reservations among patients and data owners about the formation of large collections of data and their sharing.
Data included	Rare disease patient data
Research use cases (including collaborations)	Diagnostic Registry for the Frankfurt Reference Center for Rare Diseases, registry for primary immune-deficiencies in Freiburg,

	Universitätsklinikum Ulm, Universitätsklinikum Münster, NEOCYST-Konsortium
Comments	

46. Name of service / tool	**Janus Data Repository**
Contact person	U.S. Food and Drug Administration 10903 New Hampshire Avenue Silver Spring, MD 20993 1-888-INFO-FDA (1-888-463-6332) Email: OCSServiceDesk@fda.hhs.gov
Webpage	https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm155327.htm https://patents.google.com/patent/US8484254B1/en https://en.wikipedia.org/wiki/JANUS_clinical_trial_data_repository
Country	USA
Cross-country	Yes
Short description (only a few sentences)	The Janus is a data repository for subject-level clinical and nonclinical study data submitted to FDA as part of a regulatory submission. Janus supports a transparent, consistent, and efficient scientific review process by performing automated extraction, transformation, loading, management, and integration of data to facilitate regulatory review.
Type of activity (project, service, collaboration, platform, etc.)	The Janus conceptual model is informed by the Biomedical Research Integrated Domain Group (BRIDG) Domain Model (see http://www.bridgmodel.org/).
Modules/components included	
Data included	Janus is designed to receive data in CDISC SDTM[1] format and can support other emerging study data exchange standards, such as HL7 v3, FHIR,[2] and RDF.[3] Standardized study data submitted to FDA should adhere to the study data validation rules published on the Study Data Sta [1] Clinical Data Interchange Standards Consortium, Study Data Tabulation Model; see http://www.cdisc.org [2] Health Level Seven version 3 and Fast Healthcare Interoperability Resources; see http://www.hl7.org [3] Resource Description Framework; see http://www.w3.org/RDF ndards Resources page.
Research use cases (including collaborations)	
Comments	

47. Name of service / tool	**Medical Data Space**
Contact person	Prof. Dr. Thomas Berlage Phone +49 2241 14-2141 email: thomas.berlage@fit.fraunhofer.de
Webpage	https://www.fit.fraunhofer.de/en/fb/life/medical-data-space.html
Country	Germany
Cross-country	Yes
Short description (only a few sentences)	By developing the Medical Data Space, Fraunhofer Gesellschaft hopes to make a significant, lasting contribution to reaping the benefits of digitization in the medical field through innovative IT solutions for secure exchange of medical data across institutional boundaries. In line with the aims of preventive, personalized, precise and participative medical care ("4p medicine"), this communication, which will conform to existing rules and standards, focuses on the patients.
Type of activity (project, service, collaboration, platform, etc.)	**The Initiative:** The Medical Data Space is a joint initiative of the Fraunhofer groups ICT Technology and Life Sciences. **The Concept:** The Medical Data Space (MedDS) is a virtual space that supports secure exchange and easy integration of medical and health-related data from diverse sources, using standards and shared governance models. It aims to improve the quality of diagnostics, preventive and therapeutic measures and the monitoring of therapies. MedDS protects the digital sovereignty of the data owner (patient, hospital, physician, drug company etc.) and is the basis for research and development, personalized therapies, process and cost optimization as well as new business models. MedDS helps to tap the scientific, diagnostic, therapeutic and economic potential of heterogeneous data sources.
Modules/components included	
Data included	medical and health-related data
Research use cases (including collaborations)	
Comments	

48. Name of service / tool	**The eGenVar data management system**
Contact person	T: +4793870788 E: egenvar@gmail.com
Webpage	http://bigr.medisin.ntnu.no/data/eGenVar/
Country	Norway
Cross-country	
Short description (only a few sentences)	The eGenVar provides an easy and systematic way to manage information and to advertise data. A set of tags and relationships between the donors, samples and files are used when locating information.
Type of activity (project, service, collaboration, platform, etc.)	This software is developed and maintained by the Bioinformatics and gene regulations group, faculty of medicine at the Norwegian University of Science and Technology in Norway.
Modules/components included	Java, JSP, JSON, Glassfish, Derby JavaDB
Data included	Medical associated data, Metadata
Research use cases (including collaborations)	Medical usage
Comments	

49. Name of service / tool	**ePouta IaaS Cloud**
Contact person	servicedesk@csc.fi
Webpage	https://research.csc.fi/epouta, https://research.csc.fi/cloud-computing
Country	Finland
Cross-country	Yes
Short description (only a few sentences)	ePouta is a Finnish cloud computing environment (Infrastructure as a Service, IaaS) designed for processing sensitive data. It is a closed environment that meets elevated information security level regulations. It is suitable for all fields of science, and also for government and research-sector organisations. The cloud service combines virtual computational resources with the customers' own resources using a dedicated light path. The service is easily scalable to customers' requirements.
Type of activity (project, service, collaboration, platform, etc.)	Infrastructure as a Service (IaaS)
Modules/components included	ePouta uses OpenStack technology. There is no need for separate storage capacity, as ePouta comes with storage capacity that customers can easily upgrade if required. Data is stored in Finland on CSC's seUsers can freely administer their own virtual machines and software on ePouta. The cloud resources required for each user are individualised and reserved for the user in question, and kept separate from other CSC computing environments. The customer's virtual resources are reserved in advance for a fixed period, and the customer is invoiced in accordance with agreed usage.rvers.
Data included	The computing needs can be related to scientific modeling and number crunching in various areas of science, such as life sciences, big data, astronomy, material sciences, earth sciences or financial analytics.
Research use cases (including collaborations)	cPouta Using cPouta, customers can quickly deploy self-administered servers using a simple user interface. cPouta can be used for research and educational purposes by Finnish universities and polytechnics for free. ePouta In ePouta, security is taken onto a whole new level. ePouta customers have dedicated connections from their home networks directly to their ePouta servers. ePouta is ideal for organizations that require high

	performance, high security and a certified environment for sensitive data handling.
Comments	It is not free. Base package (academic) 890 EUR Extra package (academic) 200 EUR Total cost 1090 EUR

50. Name of service / tool	**Heterogeneous Proxy Re-Encryption (H-PRE)**
Contact address / person, if available	Paper
Webpage of the tool	https://link.springer.com/article/10.1007/s11434-014-0521-1
Country it is used	When data owners store their data as plaintext in cloud, they lose the security of their cloud data due to the arbitrary accessibility, specially accessed by the un-trusted cloud. In order to protect the confidentiality of data owners' cloud data, a promising idea is to encrypt data by data owners before storing them in cloud. However, the straightforward employment of the traditional encryption algorithms cannot solve the problem well, since it is hard for data owners to manage their private keys, if they want to securely share their cloud data with others in a fine-grained manner. In this paper, we propose a fine-grained and heterogeneous proxy re-encryption (FH-PRE) system to protect the confidentiality of data owners' cloud data. By applying the FH-PRE system in cloud, data owners' cloud data can be securely stored in cloud and shared in a fine-grained manner. Moreover, the heterogeneity support makes our FH-PRE system more efficient than the previous work. Additionally, it provides the secure data sharing between two heterogeneous cloud systems, which are equipped with different cryptographic primitives.
Cross-country use?	
Short description of the tool	
Type of activity the tool supports (project, service, collaboration, platform, etc.)	
Modules / architecture / components included	
What data is stored in the tool	
Research use cases / projects / studies the tool is used (including collaborations)	We could not find medical usage.
Comments	

51. Name of service / tool	**MONOMI**
Contact address / person, if available	Paper
Webpage of the tool	https://pdos.csail.mit.edu/papers/tu-monomi-cr-vldb13.pdf
Country it is used	
Cross-country use?	
Short description of the tool	**Abstract:** MONOMI is a system for securely executing analytical workloads over sensitive data on an untrusted database server. MONOMI works by encrypting the entire database and running queries over the encrypted data. MONOMI introduces split client/server query execution, which can execute arbitrarily complex queries over encrypted data, as well as several techniques that improve performance for such workloads, including per-row precomputation, space-efficient encryption, grouped homomorphic addition, and pre-filtering. Since these optimizations are good for some queries but not others, MONOMI introduces a designer for choosing an efficient physical design at the server for a given workload, and a planner to choose an efficient execution plan for a given query at runtime. A prototype of MONOMI running on top of Postgres can execute most of the queries from the TPC-H benchmark with a median overhead of only 1.24× (ranging from 1.03× to 2.33×) compared to an un-encrypted Postgres database where a compromised server would reveal all data.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	
Modules / architecture / components included	
What data is stored in the tool	
Research use cases / projects / studies the tool is used (including collaborations)	We could not find medical usage.
Comments	

52. Name of service / tool	**Privacy-Sensitive Sharing Framework**
Contact address / person, if available	CAIDA's physical address (for drop-offs, taxis, and shuttles) is: San Diego Supercomputer Center 10100 Hopkins Drive La Jolla, CA 92093
Webpage of the tool	https://www.caida.org/publications/papers/2010/dialing_privacy_utility/dialing_privacy_utility.pdf
Country it is used	USA
Cross-country use?	
Short description of the tool	The Privacy-Sensitive Sharing (PS2) framework – that can effectively manage privacy risks that have heretofore impeded more than ad hoc or nod-&-a-wink data exchanges. Our model integrates privacy-enhancing technologies with a policy framework that applies proven and standard privacy principles and obligations of data seekers and data providers, in coordination with techniques that implement and enforce those obligations.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	Policy framework supports
Modules / architecture / components included	
What data is stored in the tool	
Research use cases / projects / studies the tool is used (including collaborations)	We could not find medical usage.
Comments	**Elements of PS2** **Authorization –** Internal authorization to share requires explicit agreement between the DP and DS. This may require direct consent from individuals identifiable in network traffic or via proxy consent with the DP.1 • **Oversight** – The DP and DS should engage external oversight of the proposed sharing, such as from an Institutional Review Board (IRB). • **Transparency** – The DP and DS should be open and in agreement over the collection, use, disclosure, objectives and obligations and associated with shared data. For example, data-sharing terms might require that the algorithms be public but that the data and/or conclusions remain protected, or vice versa. • **Compliance with applicable law(s)** – Collection, use and disclosure of data should comport to a reasonable if not case-law precedented interpretation of laws that speak directly and clearly to sharing risks about proscribed behaviors or mandated obligations.

	• **Purpose adherence** – The data should be used consistent with the documented goal for why it is being shared. • **Access limitations** – The shared data should be restricted from those who do not have a need and right to access the shared data. • **Use specification and limitation** – Unless otherwise agreed, the DP should prohibit merging or linking data that would create or enhance privacy risk. • **Collection and Disclosure Minimization** – The DP should collect and disclose only the data that is necessary to achieve the research goals, and eliminate extraneous data that carries a privacy risk. **Prominent privacy-sensitive techniques include:** o Deleting/filtering sensitive data. o Deleting/filtering part(s) of the sensitive data. o Anonymizing/hashing/de-identifying all or parts of the sensitive data. o Ag E. Mediation analysis /human proxy – this is a sandbox approach that involves "sending the code to the data" rather than releasing sensitive data for analyses. o Aging the data – traffic data that is de-sensitized by virtue of being non-current, i.e., no longer contains a direct or indirect identifier that poses a risk of harm. o Size/quantity limitation – this entails minimizing the quantity of traces shared. o Multiple layers of anonymization. • **Audit tools** – Techniques for provable compliance with policies for data use and disclosure, e.g., secure audit logging via a tamper-resistant, cryptographically protected device connected to but separate from the protected data, accounting policies to enforce access rules on protected data. • **Redress mechanisms** – Procedures to address harms from inappropriate data use or disclosure, including a feedback mechanism to support correction of datasets and/or erroneous conclusions. • **Data and analysis quality assurances** – Awareness by the DS and DP of inference confidence levels associated with the data. • **Security** – Controls should reasonably ensure that sensitive PII is protected from unauthorized collection, use, disclosure, and destruction. • **Training** – Those who are authorized to engage the data should be educated and made aware of the privacy principles and controls associated with the data.

	• **Impact assessment** – Sharing dynamics should consider potential collateral effects on stakeholders affected by the data, and seeks methods that do no further harm. • **Transfer to third parties -** This should be prohibited unless equivalent data control obligations are transferred, relative to the disclosure risks associated with that data.

53. Name of service / tool	**A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains**
Contact address / person, if available	Paper
Webpage of the tool	http://cmapspublic2.ihmc.us/rid=1P3LTQ6FM-795KWD-28CP/Sinaci_Laleci%20Erturkmen_2013_A%20federated%20semantic%20metadata%20registry%20framework%20for%20enabling%20interoperability.pdf
Country it is used	
Cross-country use?	
Short description of the tool	**abstract** In order to enable secondary use of Electronic Health Records (EHRs) by bridging the interoperability gap between clinical care and research domains, in this paper, a unified methodology and the supporting framework is introduced which brings together the power of metadata registries (MDR) and semantic web technologies. We introduce a federated semantic metadata registry framework by extending the ISO/IEC 11179 standard, and enable integration of data element registries through Linked Open Data (LOD) principles where each Common Data Element (CDE) can be uniquely referenced, queried and processed to enable the syntactic and semantic interoperability. Each CDE and their components are maintained as LOD resources enabling semantic links with other CDEs, terminology systems and with implementation dependent content models; hence facilitating semantic search, much effective reuse and semantic interoperability across different application domains. There are several important efforts addressing the semantic interoperability in healthcare domain such as IHE DEX profile proposal, CDISC SHARE and CDISC2RDF. Our architecture complements these by providing a framework to interlink existing data element registries and repositories for multiplying their potential for semantic interoperability to a greater extent. Open source implementation of the federated semantic MDR framework presented in this paper is the core of the semantic interoperability layer of the SALUS project which enables the execution of the post marketing safety analysis studies on top of existing EHR systems.
Type of activity the tool supports (project, service, collaboration, platform, etc.)	
Modules / architecture / components included	CDISC SHARE, HITSP, OMOP, Semantic MDR
What data is stored in the tool	metadata

Research use cases / projects / studies the tool is used (including collaborations)	
Comments	

54. Name of service / tool	**Efficient Security Framework for Sensitive Data Sharing and Privacy Preserving on Big-Data and Cloud Platforms**
Contact person	Paper
Webpage	https://dl.acm.org/citation.cfm?id=2896423
Country	
Cross-country	
Short description (only a few sentences)	Cloud computing storage was widely used for storing user's data, but cloud computing only providing the tasks of data storage, but not supporting the important functionalities like computation and database operations. These operations are supported by big data systems and hence currently use of big data platform for storage in increases worldwide by enterprises. Sharing sensitive information and data resulted into big reduction in costs of enterprises for users to provide value added data and personalized services. As enterprises are sharing their important and sensitive information on big data, it becomes necessary to provide the security and privacy in big data platform. For Big Data platforms, secure sharing of sensitive data is challenging research problem. In this paper, different security and privacy preserving methods of cloud computing and big data platforms are introduced, and then a novel hybrid framework for secure sensitive data sharing and privacy preserving public auditing for shared data over big data systems is presented including functionalities such as privacy preserving public auditing, data security, storage, data access, deletion or secure data destruction using cloud services.
Type of activity (project, service, collaboration, platform, etc.)	
Modules/components included	
Data included	
Research use cases (including collaborations)	
Comments	

55. Name of service / tool	**Fast Healthcare Interoperability Resources (FHIR)**
Contact person	Standard
Webpage	hl7.org/fhir/

Country	International HL7 organisation
Cross-country	international
Short description (only a few sentences)	It is a standard describing data formats and elements (resources) and an API for exchanging EHR data. It builds upon HL7 v2 and v3 and the data format is JSON.
Type of activity (project, service, collaboration, platform, etc.)	Data exchange standard
Modules/components included	
Data included	A test server with a representative set of data has been made available at: https://try.smilecdr.com:8001/
Research use cases (including collaborations)	The benefits of FHIR for clinical research within the biopharmaceutical community is obvious; FHIR resources can be used as eSource data to pre-populate (eCRFs) clinical research case report forms for both regulated and non-regulated clinical trials. Connectathons have been realised with the participation of pharmaceutical companies, members of TransCelerate Biopharma, Inc. and technology implementers to simulate using FHIR to populate and manage clinical study databases. For example it is possible to use of HL7 FHIR as eSource to Pre-populate CDASH CRFs using a CDISC ODM API. Another use case is the use of FHIR for the generation of Real World Evidence (RWE) for clinical trials.
Comments	As stated in its May 2016 FDA draft guidance “Use of Electronic Health Record Data in Clinical Investigations”: the FDA encourages sponsors and clinical investigators to work with the entities that control the EHRs, such as health care organizations, to use EHRs and EDC systems that are interoperable establishing interoperability between EHR and EDC systems to streamline and modernize clinical investigations for improving data accuracy, patient safety, and clinical research efficiency.