Long Tail of Data

e-IRG Task Force Report

September 2016

http://e-irg.eu

Colophon

Acknowledgments

The e-IRG Task Force on the Long Tail of Data created and edited this document with the help of the community.

For the content: Françoise Genova and Wolfram Horstmann

This document was produced by the e-IRGSP4 project for e-IRG.

e-IRGSP4 is supported by the FP7 Capacities Programme under contract nr RI-632688.

This work is licensed under a Creative Commons Attribution 4.0 International License.

ISBN 978-90-823661-3-6

Version 1.74

e-IRG secretariat

P.O. Box 93575,
NL-2509 AN The Hague,
The Netherlands
Phone: +31 (0)70 344 0526
E-mail: secretariat@e-irg .eu

Visiting address

Java Building, Laan van Nieuw Oost-Indië 300
NL-2593 CE The Hague
The Netherlands

Table of Contents

1 Executive Summary

Data driven science is seen as the fourth paradigm[1] of scientific research after experimental science, theoretical science and computational science. With the advent of Big Data the focus turned towards massive data volumes and the very large amount of small data sets became out of sight of funders, policy makers and the general public. Today it is time to put attention back to the Long Tail of science, be it for data, computing or services, and bestow it the recognition, which it deserves. In the movement towards Open Science and cross-disciplinary research the enormous value of small and medium data sets is recognized. Thus the requirements of the Long Tail of science should be taken into account during the implementation of international, national and local e-Infrastructures and furthermore the Long Tail of science should benefit of all developments that take place to solve the e-Infrastructure requirements of researchers.

This document focuses on the Long Tail of scientific data, its sustained management and storage. Its characteristics are discussed, and it is shown that the distinction between the Long Tail and Big Data is somehow blurred. Long tail data management should sustain trust in data and repositories, with as keywords data quality, certification of repositories, appraisal, documentation, discoverability and interoperability. The design principles of data repositories are of crucial importance and will have long-term impact, so that it is of critical importance to adapt standards, access policies, governance rules, the integration of industry, etc. to both, Big Data as well as the Long Tail of data. Policy actions should be adapted to the characteristics of Long Tail data and of their producers and users. More work is needed to understand how different building blocks can intervene to build an operational, relevant ecosystem. This has to be taken into account especially in the realization of the European Open Science Cloud.

Recommendations to improve the situation of Long Tail data are are expressed on different relevant aspects. On Policies, funders and institutions should define, publish and disseminate their policies, privileged open licences , and the certification of repositories encouraged. Data should be shown in all occasions by scientists and institutions as a scientific, institutional and societal asset, and publications should be linked to the relevant data. As Incentives, it is important to gather success stories, to show researchers that it is easy and beneficial to deposit their data, and to illustrate problems of irreproducibility of research, duplicated researches and innovation loss. Among the key technical building blocks identified for improving access to and usage of datasets, it is recommended to push

[1] *The Fourth Paradigm*, edited by Tony Hey, Stewart Tansley, and Kristin Tolle, http://www.fh-potsdam.de/fileadmin/user_upload/fb-informationswissenschaften/bilder/forschung/tagung/isi_2010/isi_programm/TonyHey_-__eScience_Potsdam__Mar2010____complete_.pdf

standards and technologies across disciplines, to implement persistent identifiers for all elements of the system, to increase discoverability and to implement a discovery layer in the form of a landing page. In general, all policy actions should be adapted to the characteristics and needs of the Long Tail, and more work is needed to understand how to build a relevant, operational ecosystem.

2 Introduction - "Big" and "Long Tail" Data

In the Open Science arena attention is currently focused on the Big Data phenomenon but there is other data that falls out of this focus. *Data diversity* can be regarded as a more general organizational principle of research. On one side of the data diversity spectrum, Big Data addresses the exponential growth of data generation and availability, with its multiple opportunities and challenges for scholarship (Boyd & Crawford, 2012). Big Data comprises structured as well as unstructured data with a tendency to be homogeneous and standardized (Borgman, 2015). On the opposite side of that spectrum, a Long Tail has been described to stress the variety in structure, subject, complexity, context, format, size, location, and its use in research (Heidorn, 2008).

In statistics, the Long Tail is a specific form of a probability distribution in which a small part of a given distribution has many occurrences (the head) while a large part of the population has comparatively few occurrences (the tail). In research, the Long Tail of data can be characterized as heterogeneous, relatively small data that have unique standards and are not regulated – implying the necessity of personalized curation and control in smaller institutional and domain repositories (Heidorn, 2008). This description of the Long Tail is instructive to understand the diversity of data in a general sense. But there are several problems with the Long Tail metaphor.

"Long Tail" implies that it deals with the rest of data, the not-so-important parts or the small parts that "do not" need attention. This 'belittling' notion of the Long Tail is challenged even by major research areas. For the Brain Sciences, for example, it was pointed out that the Long Tail may in fact represent the mainstream research (Ferguson, Nielson, Cragin, Bandrowski, & Martone, 2014). Another example is Astronomy, in which the world-wide, distributed and interoperable data infrastructure contains the observations of the ground- and space based telescopes, which are made available in the observatory archives and belong to the Big Data side. But it also contains data that belongs to the Long Tail side, in particular research results attached to publications in the academic journals, curated in a disciplinary data centre, the CDS (Centre de Données astronomiques de Strasbourg[2]). One can also cite another specialized repository, PANGEA[3], which hosts and disseminates geo-localized data and also works in collaboration with journals. A second problem of the Long Tail metaphor is arising from the simplicity of the underlying statistical concept: it indicates that there is only one dimension along which diversity exists – for example it can be either size or format but not both. For properly characterizing the research data landscape, however, a multidimensional approach is required (see also Borgman, 2015). In addition, Big Data can be diverse and not well structured, and thus, share the main characteristics of Long Tail Data.

[2] http://cds.unistra.fr/
[3] https://www.pangaea.de/

3 Characteristics and Challenges of Long Tail Data

A rough summary comparison of Big and Long Tail Data, adapted from Heidorn (2008), could be the following, even if in reality, as explained, the situation can be much more fuzzy:

Nr.	Head	Tail
1	Homogeneous	Heterogeneous
2	Large	Small
3	Common standards	Unique standards
4	Regulated	Not regulated
5	Central curation	Individual curation
6	Disciplinary repositories	Institutional, general or no repository

Long tail data exist across all disciplines, very often only in individual computers or university servers, and often also with minimal or no attached metadata or documentation, a major obstacle to reusability. Many of the characteristics of data are discipline dependent. Heterogeneity/diversity has many dimensions, among which:

- Size: research data can vary by dozens of orders of magnitude, from a few bytes in an ASCII table to Exabyte scale in large research facilities;
- Format: this includes the diversity in the original formats, with the frequent usage of proprietary and/or disciplinary formats, compression and aggregation;
- Structure: the capacity to refer to an explicit or implicit data model is by far not guaranteed;
- Complexity: data can be composed of different elements, can have different versions or vary with time.

The Long Tail of data is thus often in trouble, but it is ubiquitous and taking it properly into account is one of the keys to optimally exploit the capacity of Open Science. It raises more and more interest, in particular with the development of institutional and general repositories, including the public and commercial ones, and in the library community, since librarians undergo a very significant evolution of their traditional tasks and are interested in becoming scientific data curators. Another incentive is the fact that research funders more and more require the projects they fund to develop a Data Management Plan: researchers applying for funding need to think about the fate of at least some of their Long Tail data.

The Long Tail is a meeting point of the research, library and IT communities, as it appears when looking at some of the current relevant activities at the international and European levels:

- The Research Data Alliance *Long Tail of Research Data Interest Group*[4] (IG): The aim of the IG founded in 2013 was to develop a set of good practices for managing research data archived in the university context. The scope is focusing on the data generated in universities and research institutions and on the role of repositories and libraries as agents of the institutional data management. The IG document shows how repositories tackle Long Tail data and the challenges they encounter;
- Universities: The League of European Research Universities (LERU) has produced a Roadmap for Research Data[5] in 2013 that is taking an institutional perspective expressly stressing the role of the Long Tail of research data. Recently, the European University Association (EUA) has founded a Science 2.0 / Open Science[6] group that takes a university perspective;
- Libraries: Diverse, institutionally produced data naturally lie in the interest of libraries, since they are the locally responsible long-term institution for information provision. The European Library Federation LIBER released the Ten Recommendations for Libraries to get started with Research Data in 2012[7] and produced case studies[8] for service implementations. Libraries also form one of the largest interest groups in RDA[9] which mediates with the global community and has produced a series of briefing papers and recommendations;
- Journals: A major driver in the domain of Long Tail data are journals that increasingly request data sets underpinning the research findings in scientific articles to support reproducibility of research and prevent scientific fraud. Obviously, these requests by journals drive researchers to find appropriate repositories for these specific datasets that form part of the Long Tail. An overview of repositories is provided by Re3Data[10], which lists almost 1500 repositories as of 2016. Additionally, specific journals are being launched for data intensive research such as F1000 Research[11] or specific 'data papers' such as Nature's Scientific Data[12]. Publishers also provide data sharing platforms such as Figshare[13];
- e-Infrastructures: The Open Access Infrastructure for Research in Europe (OpenAIRE) has started with a focus on implementing EC policies on research literature but addressed already in its precursor project DRIVER in 2007 combinations of literature and data

[4] https://rd-alliance.org/groups/Long Tail-research-data-ig.html
[5] http://www.leru.org/index.php/public/news/press-release-leru-roadmap-for-research-data/
[6] http://www.eua.be/policy-representation/research-innovation-policy/science-2-0-open-science
[7] http://libereurope.eu/blog/2012/08/24/ten-recommendations-for-libraries-to-get-started-with-research-data-management/
[8] http://libereurope.eu/committees/scholarly-research/research-data-management-case-studies/
[9] https://rd-alliance.org/group/libraries-research-data/post/libraries-research-data-interest-group-endorsed-rda.html
[10] http://service.re3data.org/search
[11] http://f1000research.com
[12] http://www.nature.com/sdata/
[13] https://figshare.com

(Enhanced Publications[14]), which recently culminated in Data-Literature-Interlinking Service[15] jointly produced by OpenAIRE (in the context of RDA) in collaboration with the World Data System. A major service that has been established in the context of OpenAIRE is the repository Zenodo[16] which also hosts numerous datasets. OpenAIRE is recently also focusing on the EU policies in relation to the Open Research Data Pilot[17]. EUDAT has started around 2011 and is addressing the Long Tail, for example, with specific data pilots[18].

Establishing an efficient management of Long Tail data is thus required but is challenging because of its current status and huge diversity. The challenges include technical and sociological aspects. As explained e.g. by the G8 Ministers of Research in their strong June 2013 statement about scientific data[19], "open scientific research data should be easily discoverable, accessible, intelligible, useable, and wherever possible interoperable to specific quality standards" – some of these principles 'Findable/Accessible/Interoperable/Reusable' constitute the FAIR principles[20]. The elements that are part of Trust in data (data quality, certification, appraisal, documentation, discoverability, interoperability) are the same for head data and for Long Tail data but are less evident for the latter category. Trust is an essential element of scientific data sharing. Data quality is one of the key points, and one of the basis of Trust. Another element of Trust for the data provider and the users is confidence in the ability of the repository to manage the data properly on the long term, which can be ensured through an external Certification process. Appraisal is a key component of Long Tail data management – data preservation and dissemination has a cost and not all data will be preserved. The data to be preserved must have enough documentation (metadata, provenance, etc.) to be intelligible and reusable. Discoverability is another one and one important strand of work is to increase discoverability in diverse repositories. Interoperability is also a must since by essence the data will be stored in distributed repositories.

[14] http://dare.uva.nl/cgi/arno/show.cgi?fid=150723
[15] http://dliservice.research-infrastructures.eu/#/
[16] https://zenodo.org
[17] https://www.openaire.eu/opendatapilot
[18] http://eudat.eu/eudat-call-data-pilots
[19] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf , Section 3 Open Scientific Data
[20] https://www.force11.org/node/6062

4 Repositories for the Long Tail

In this domain in particular, the respective roles of institutional, generic and disciplinary registries should be ascertained. This is linked to the type of metadata (e.g. basic Dublin Core plus in addition disciplinary vocabularies/ontologies for discoverability and usage) and to the capacity to do quality assessment and checks of the data and metadata, and to provide APIs relevant to scientific needs.

The research data infrastructures must take fully into account providers' and users' requirements. Long tail data (as well as bigger data!) can be made available at a basic level allowing discovery by common metadata such as author or curator, very similar to those which are used for scientific publications. Including the disciplinary knowledge in particular in metadata allows advanced, fine-tuned discovery and usage. For instance, geo-localised data should be discoverable through queries by space coordinates.

The recent discussions about Big Data and the emergence of cloud technology give rise to visions of comprehensive e-Infrastructures that shall benefit a cross-national and trans-disciplinary research community. The success of these visions critically depends on its aptitude to research, i.e. in how far it implements imperatives of research and how far it supports data diversity. Similar initiatives already have proven prone to risks of disruptive innovation, notably the development of grid and cloud technology (EUDAT, LIBER, OpenAIRE, EGI, GEANT). Great care should be taken to take a holistic stance and respect local research practice in order to ensure that the design and construction of the data infrastructure is fully user-driven and not dominated by technology or 'political' aspects.

Which design principles can be deducted from the characteristics of Long Tail data? There are many ways how research data infrastructure can be designed. For the sake of clarity, three simplified scenarios shall be considered here:

1. Industry-Owned. In this scenario, infrastructure is built in a private-public partnership. Industry deploys capable e-Infrastructure for its own purposes and fierce competition makes it robust and cost-effective. Risks include the loss of control and ownership, the application of proprietary standards and restrictive licenses as well as prevalence of commercial interests. This scenario thus shows weak support of research imperatives, specifically communalism and skepticism. But the scenario is possible in principle, if risks were controlled;
2. Publicly-Owned. In this scenario, infrastructure is built and operated by a public entity, say an international NGO. Since commercial interest can be excluded, it can directly foresee a data

commons by implementing Open Access policies. However, it runs the risk to become too generic and less responsive to specialized community needs and to end up in a mediocre applicability serving the least common denominator, thereby oversimplifying and restricting innovation capacity;

3. Community-Owned. In this scenario, specialized communities receive funding for building their own e-Infrastructure, making sure that their requirements are met. Control for development is led by specialized needs and innovation capacity is maximized, thereby reducing risks of central and externally owned solutions. But this scenario also introduces risks of redundancy and transdisciplinary barriers by building specialized silos.

The three scenarios described above reveal a complex matrix of benefits and risks. Of course, neither of the three scenarios in its pure form is likely to occur. It is recommended here to apply a 'best-of-breed' approach for research data infrastructure that is adhering to specific design principles, establishing a smart infrastructure for Long Tail (and other) data:

- Research data infrastructure should be rigorously analyzed whether it support research values, i.e. academic freedom, universalism and data diversity;
- Open standards should be used in a rigorous but parsimonious way – not to restrict expressiveness of research data and keeping control of the ways data are processed and shared;
- Proprietary standards and restrictive licenses for data access must require justification,
- Industry services should be used if (and only if) it is beneficial to the effective conduct of research;
- Research data infrastructure must not interfere with requirements of researchers, i.e. the degrees of freedom of applying infrastructure must be maximized and controlled by the researchers; centralized control must be minimized;
- Governance of research data infrastructure must implement control of diverse subject communities and methodological cultures, be democratic – any central governance must be prevented.

5 The European Open Science Cloud and the Long Tail of Data

In the motion of the European Parliament "Towards a Digital Single Market Act", it is stated: "124. Is concerned that cloud infrastructures for researchers and universities are fragmented; calls on the Commission, in cooperation with all relevant stakeholders, to set up an action plan to lead to the establishment of the European Open Science Cloud (EOSC) by the end of 2016, which should seamlessly integrate existing networks, data and high-performance computing systems and e-Infrastructure services across scientific fields, .."[21]

Hence, the European Cloud Initiative[22] has been started. The communication sets the objective "to ensure that European scientists reap the full benefits of data-driven science. Practically, it will offer 1.7 million European researchers and 70 million professionals in science and technology a virtual environment..." and "...providing easier access via the Cloud both to researchers in key scientific disciplines and to the Long Tail of science."

While the first statement shows an inclusive and participatory approach, the second statement differentiates "key scientific disciplines" and the "Long Tail of science" which implies an opposition between the two. It is not clear how "key disciplines" are defined but "key areas" mentioned elsewhere in the text are health, environment and transport, which may mean that the "Long Tail of science" here includes all the so-called 'knowledge driven' sciences, irrespective of the fact that they mainly use big or small data. It has been shown that some of the mentioned "key scientific disciplines" actually base their work on the Long Tail of research data. Neuroscience researchers, for example, stated in a Nature publication[23] that Long Tail data is the essential prerequisite for producing excellent research. Many outcomes of the Copernicus programme will require that satellite and in-situ data, often obtained by specific targeted programmes, be used together.

If the EOSC is focussing on 'key players' with Big Data and not including the diversity of research in Europe, the research infrastructure might be built in a way that it fails to support the excellence in European research for two reasons:

- Scientific innovation is often generated at the vertices between disciplines not within one big discipline as the rise of cross-disciplinary research in international networks in the last two

[21] http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A8-2015-0371+0+DOC+XML+V0//EN
[22] http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52016DC0178&from=EN
[23] http://www.nature.com/neuro/journal/v17/n11/full/nn.3838.html

decades has demonstrated. Focussing on ”key areas“ when building the EOSC might mean to miss the requirements of actual innovation drivers in European research.

- The breadth and excellence of European research lies in its distributed diversity of universities and research institutions. A forte of cloud infrastructure is that seamlessness and cost-efficient central structures can be combined with customized adaptations to local needs. A strong focus on key disciplines, however, might lead to a situation in which most of the investment goes into central processes with some selected adaptations for key areas and only a small fraction supports the diversity where innovation and excellence has its basis.

In summary, for supporting excellence and innovation for the 1,7 million European researchers, it is suggested here that the main recommendation for building the EOSC is to support diversity. Taking properly into account the Long Tail of data will be a key indicator of the EOSC final success, since it will demonstrate its relevance and its capacity to fully integrate the local and disciplinary levels.

6 Recommendations

Policies

- Funders and institutions should define their policies;
- Open licences to be privileged;
- Encourage certification of repositories;
- Scientists/institutions: Show data as scientific/institutional/societal asset;
- Link data to publications (generic or disciplinary endeavor).

Incentives

- Funders and institutions should publish and disseminate their policies;
- Gather success stories (institutions/RDA Group);
- Show researchers how easy and beneficial it is to deposit data;
- Show problems of irreproducibility, double research and innovation loss.

"Technical building blocks" for improving access to and usage of datasets

- Push standards and technologies across disciplines;
- Persistent identifiers for data, institutions, people e.g. DataCite-DOIs, ORCID;
- Discoverability: increase discoverability in diverse repositories;
    - Dataset and repository registries;
    - Link data to publication;
    - Discovery layer – landing page,

More generally, policy actions should be adapted to the characteristics of Long Tail data and of their producers and users. More work is needed to understand how different building blocks can intervene to build an operational, relevant ecosystem. Some of the recommendations could be suggestions for the RDA Interest Group or other existing or possible RDA Groups.

7 References

Borgman, C. L. (2015). Big data, little data, no data: scholarship in the networked world.

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662–679. http://doi.org/10.1080/1369118X.2012.678878

EUDAT, LIBER, OpenAIRE, EGI, & GEANT. (2015). Position Paper: European Open Science Cloud for Research. http://doi.org/10.5281/zenodo.32915

Ferguson, A., Nielson, J., Cragin, M., Bandrowski, A., & Martone, M. (2014). Big data from small data: data-sharing in the "Long Tail" of neuroscience. Nat. Neurosci., 17(11), 1442–1447. http://doi.org/10.1038/nn.3838

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. Library Trends, 57(2), 280–299.

8 Glossary

Term	Explanation
ASCII	American Standard Code for Information Interchange is a code for representing English characters as numbers
API	Application Programming Interface
CDS	Centre de Donnés astronomiques de Strasbourg
DataCite	DataCite is a leading global non-profit organisation that provides persistent identifiers (DOIs) for research data.
DOI	Document Identifier
DRIVER	Digital Repositories Infrastructure Vision for European Research
e-IRG	e-Infrastructure Reflection Group
EC	European Commission
EGI	A federation of resource centres and coordinated by EGI.eu
EOSC	European Open Science Cloud
ESFRI	European Strategy Forum on Research Infrastructures
EU	European Union
EUA	European University Association
EUDAT	European Data Infrastructure
G8	Group of 8 (now G7 due to Russia'a suspension) is a governmental political forum of the leading industry nations and the EU
GÉANT	Pan-European research and education network that interconnects Europe's National Research and Education Networks (NRENs).
IT	Information Technology
IG	Interest Group
LERU	League of European Research Universities
LIBER	Ligue des Bibliothéques Européennes de Recherche (Association of European

	Research Libraries)
NGO	Non-governmental Organisation
OpenAIRE	Open Access Infrastructure for Research in Europe
ORCID	ORCID is a not-for-profit organization that provides an identifier for individuals to use with their name as they engage in research, scholarship, and innovation activities.
PANGEA	Data Publisher for Earth & Environmental Science is a digital data library and a data publisher for earth system science
RDA	Research Data Alliance
Re3Data	re3data.org is a global registry of research data repositories that covers research data repositories from different academic disciplines

http://e-irg.eu