"You say potato, I say potato" - Mapping Digital Preservation and Research Data Management Concepts towards Collective Curation and Preservation Strategies

Michelle Lindlar
TIB - Leibniz Information Centre for Science and Technology

Pia Rudnik
ZB MED - Information Centre for Life Sciences

Sarah Jones
Digital Curation Centre

Laurence Horton
University of Toronto, School of Information Studies

Abstract

This paper will explore models, concepts and terminology used in the Research Data Management and Digital Preservation communities. In doing so we expect to identify several overlaps and mutual concerns where the advancements of one professional field can apply to and assist another. By focusing on what unites rather than divides us, and by adopting a more holistic approach we will advance towards collective curation and preservation strategies.

Introduction

Both research data management (RDM) and digital preservation (DP) communities take on responsibility for keeping research data reusable. Since digital research data is an information type of archival value, the two communities are, in theory, moving closer together. Conferences like the International Digital Curation Conference (IDCC) or the International Conference on Digital Preservation (iPRES) are a perfect example of where the two communities come together - but do they really collaborate or are they only co-existing? Classical DP topics such as file format validation, preservation policies or web archiving are traditionally under-represented at IDCC. Along similar lines, RDM topics, such as Data Management Planning tools, or the engagement of the scientific community in creating metadata needed for curation, are underrepresented at iPRES, the premier conference for digital preservation. Is this lack of interdisciplinarity at the two conferences symptomatic for the way the DP and RDM communities interact at large?

RDM is typically focused on the creation process of research data by making data producers accountable for creating and maintaining well-documented data in sustainable

Submitted 08 July 2019

Correspondence should be addressed to Michelle Lindlar, Technische Informationsbibliothek (TIB) German National Library of Science and Technology, Digital Preservation, Welfengarten 1 B // 30167 Hannover, Germany Email: michelle.lindlar@tib.eu

form so it can be curated throughout the entire lifecycle of data. DP comes into play when a research data set is deemed to have archival value or when the lifecycle exceeds a short-term timeframe. But how can we define a "short-term timeframe" and is it really early enough to think in DP terms after that timeframe has expired?

Over the course of the last decade, RDM has defined itself to operate within a timeframe of at least ten years. The Deutsche Forschungsgemeinschaft (2018) has a minimum ten year preservation requirement, whereas the UK Engineering and Physical Sciences Research Council's (EPSRC 2015) expectation is to preserve data for ten years after the last request for use. Some institutions have their own specified minimum retention period. Other research funders and publishers have open ended expectations for keeping data available and recommend a data repository to ensure long-term access. Example are Canada's Tri-Agency Research Data Management Policy (Government of Canada 2019) or Springer Nature's Research Data Policy (2019).

Requirements in repository certification foresee a handover point, where research data of archival value is transferred to an archive to ensure long-term availability. The German Initiative for Network Information (DINI) Certificate for Open Access Repositories and Publication Services (Müller, Scholze et al., 2016) explicitly addresses the hand-over between mid- and long-term availability in its "Long-Term Availability" criteria. Certified repositories are required to keep documents and metadata published available for "a minimum time span of no less than five years". The certificate furthermore recommends that long-term availability is ensured through a cooperation between the certified repository and a DIN 31664 "Information and Documentation Criteria for Trustworthy Digital Long-Term Archives" certified archiving institution.

However, a handover of data between RDM and DP at a fixed point in time bears risks. Problems such as corrupt files, subpar file formats, insufficient metadata, and missing provenance and rights information might exist early on in the lifecycle of research data. This can lead to a resource-intensive DP process or even make preservation impossible. To us, this could be avoided by achieving synergies between RDM and DP good practice and by fostering collaboration between content creators and curators. In our view, collaboration begins with a shared understanding of core concepts and the terminology used in both the RDM and DP community. This leads us to the question, which key terminology do the two communities use in these core concepts. Do we really speak the same language and follow the same goals?

Research Question

This paper will take an in-depth look at core concepts used by the RDM and DP communities, checking whether they support a shared understanding of models and processes. The authors examine whether the identified core concepts can be interlinked, mapping how RDM and DP concepts intersect, paving the way towards a collective curation and digital preservation process.

Methodology

The analysed core concepts are what we consider to be the main models, (de-facto) standards, and process descriptions within one or both of the communities. The following core concepts will be analysed against the aforementioned research question:

- DCC curation lifecycle model
- Object Levels of Preservation (according to the works of Kenneth Thibodeau)
- Data Management Plans
- FAIR
- OAIS
- PREMIS

Key terminology has been extracted from each model and transferred to a table in Appendix A. This table represents a mapping generated via discussion between the authors, who consist of two RDM domain experts and two DP domain experts. Within each section of the paper, key terminology used in the mapping is emboldened. Alternate terms applied in use cases are italicised.

As mutual understanding of terminology needs to be built on a strong foundation. There are some overarching key terms we felt it necessary to define as context for the paper. These are the two communities (Research Data Management and Digital Preservation) and the difference between archive and repository. For the purposes of our paper, they are differentiated as follows:

- Research Data Management is most concerned with activities which happen early in the lifecycle of a digital object. In our understanding of the term, it is mainly a producer and process oriented activity, including tasks such as creation, selection and enhancement of data. Consumers in the scope of RDM are considered in a short- to mid-term timeframe.
- Digital Preservation is a set of formal actions focused on ensuring long-term availability and interpretability - from a technological as well as semantic aspect - of all data (publications, research data, metadata, etc.). RDM can be regarded as a first step towards preservation, as important information, e.g., about the creating process and intent needs to be captured at this stage. In contrast to the rather process-oriented RDM, DP is more object-oriented. DP processes mainly focus on the requirements of the Consumers in the scope of mid- to long-term timeframes.

A second differentiation in terminology which was discussed in detail by the authors was that between “repository” and “archive”. In contemporary literature they are often used synonymously, however for the purposes of this paper, we use archive in the OAIS sense of the word, as an “organization, which may be part of a larger organization, (consisting) of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community”. An archive is therefore preservation focused. A repository may be an organization and combination of software and people in the same sense an archive is, however, a repository offers a data service. Repositories are therefore not necessarily preservation focused. Many repositories focus on providing access to content and are not actively involved in digital preservation.

Core Concept Analysis

For each concept we give a brief description, extract key terminology[1] and demonstrate applications via use cases. A general critique is provided to examine adoption and suggest wider applications or lessons that can be learned from the model to bring the DP and RDM communities closer together.

DCC Lifecycle Model

Presenting a view that encompasses both, typical RDM and typical DP tasks, the DCC lifecycle model is a natural candidate for a model to consider within the scope of this paper. The Curation Lifecycle Model (Higgins, 2008) was developed by the Digital Curation Centre (DCC) to articulate the key activities involved in managing content from the initial planning phases to long-term reuse. Curation and preservation are understood as interlinked concerns encompassing activities undertaken by content creators (e.g. planning and data creation) as well as digital preservation processes (e.g. preservation planning, migration) and interactions with external stakeholders (e.g. access & reuse and community watch and participation). Broad terminology (such as digital objects) was applied to enable the lifecycle model to be applied to all types of content and contexts. Work has recently been undertaken by organisations in the United States and Korea to rethink and update the lifecycle model, providing new applications and meanings (Sveinsdottir, forthcoming)

Key Terminology of the Model

At the core of the digital curation lifecycle model (see Figure 1) is the data (**digital objects**), together with the associated description and metadata (**representation information**) to make the data meaningful. The rest of the model is broken down into sequential actions (the burgundy boxes) or continuous activities (the inner circles). The primary emphasis of the model is on the interlinked activities to curate and preserve content throughout its lifecycle. The DCC lifecycle model includes activities that fall in the content curator role such as conceptualisation (where DMPs or project proposals would be authored), **create** and **appraise & select,** as well as task such as **ingest** and **preservation planning**, **preservation action / transform** and **community watch** which have clearer application in digital preservation communities. The lifecycle model anticipates that different stakeholders will be responsible for different elements and encourages collaboration across the groups. The DCC lifecycle model could be considered environment agnostic, as it does not explicitly mention actors such as data producer, archive and consumer, instead only describing the functions conducted by these actors, i.e., **create / receive**, **store** and **access**.

[1] Extracted terminology which we included in a mapping presented in the conclusion of this work is indicated in bold type throughout the chapters

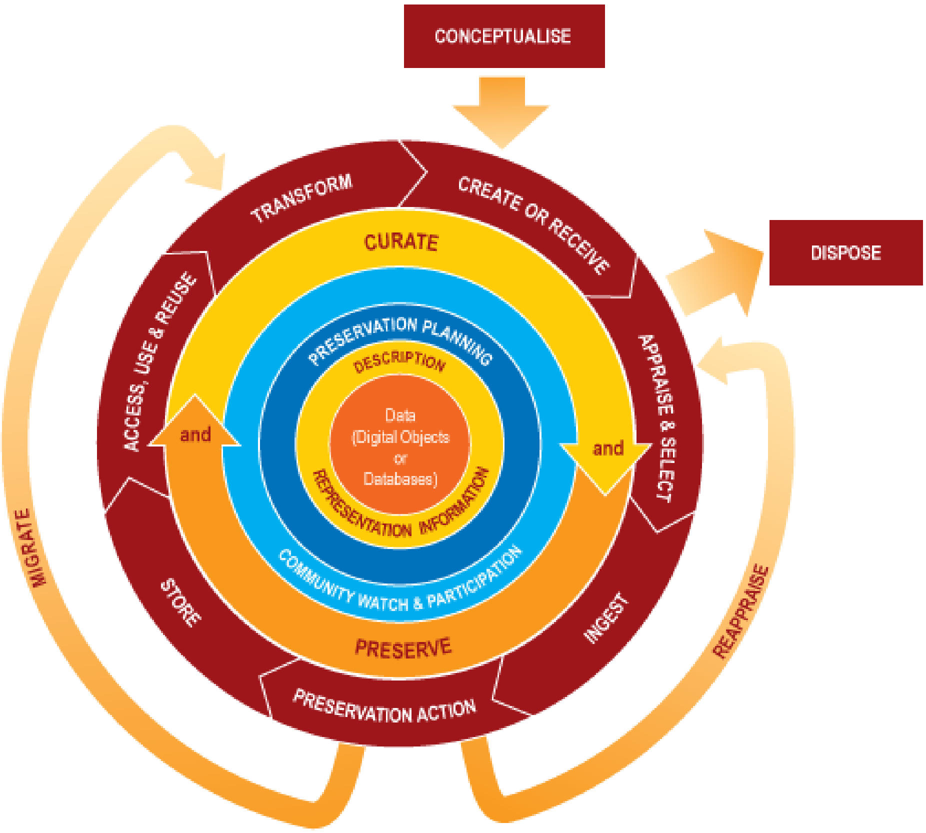


Figure 1. : DCC Lifecycle Model (Higgins, 2008).

Use Case - implementing the lifecycle model in the KISTI Content Curation Centre

The Korean Institute of Science and Technology Information (KISTI) have been implementing the lifecycle model in the Content Curation Centre. This team is tasked with collecting all Korean research reports and articles by harvesting metadata from data centres and journals or holding copies of research papers not available elsewhere. Some key changes were made to the model during implementation. The KISTI Model renames the Description and Representation Information circle as *Semantic Description* a term which is likely to be more generally understood than **Representation Information** (see also OAIS). An additional full lifecycle action of *Enhancement* has been added to emphasise that curation actions should improve the data for example by providing additional metadata and description, assigning identifiers, converting to new formats, or linking to external resources. **Community Watch** has been amended to *Stakeholder Observation* to emphasise the range of stakeholders that contribution to the curation of data and the need to engage with them. The final addition was that of *User Experience* to register users and capture data about their interactions. The DCC is keen to explore whether *User Experience* could be integrated into the DCC model and broaden this out to include also more qualitative approaches to capturing user feedback with the specific aim of improving RDM services.

General Adoption and Critique

The DCC Lifecycle Model has become a canonical resource, referenced broadly in the digital curation, preservation and research data management communities. The recent work to iterate on the model and apply it to new contexts demonstrates its validity more than a decade after it was first created. In terms of RDM, the model is often too complex to demonstrate to researchers, especially as so many actions focus on the role of the curator. Research support staff tend to use research focused lifecycle models that emphasise steps such as data analysis, storage and publication.

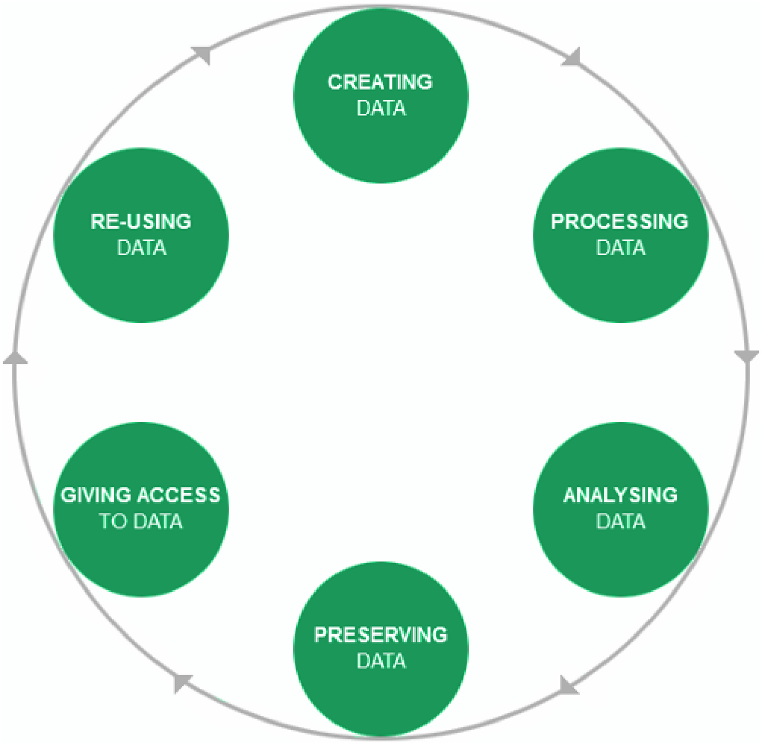


Figure 2: Research Data Lifecycle by the UK Data Archive

Object Levels of Preservation

While the DCC Lifecycle Model clearly focuses on the process, a core digital preservation model - the Object Levels of Preservation - focus on the digital object itself. In 2002, the archivist Kenneth Thibodeau gave an “Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years” and described a holistic approach to digital preservation that focuses on the digital objects of archival value themselves (Thibodeau, 2002). In this paper, we refer to Thibodeau’s model as “Object Levels of Preservation”, not to be confused with the Levels of Preservation developed by the National Digital Stewardship Alliance. Thibodeau’s model is based on the definition of three digital object levels (physical, logical, and conceptual), which together need to be considered in DP.

Key Terminology of the Model

Thibodeau’s model defines three different preservation levels in accordance to the main characteristics of digital objects. **Digital Object** is generally defined as “an information object, of any type of information or any format, that is expressed in digital form.” Regarding the basic characteristic of digital objects, Thibodeau points out: “All digital objects are entities with multiple inheritance [...], the properties of any digital object are inherited from three classes.” These classes are the **Physical, Logical**, and the **Conceptual**

Object, which themselves have unique properties (see Figure 2) and bear distinct risks, e.g. lack of robust storage on the physical level, unsuitable file formats or dependencies on externally linked resources on the logical level, and incomplete accompanying descriptive metadata or semantic drift on the conceptual level. In describing different object levels, their properties and corresponding risks, Thibodeau's model emphasizes the need for different preservation methods that correspond to these levels and mitigate the risks.

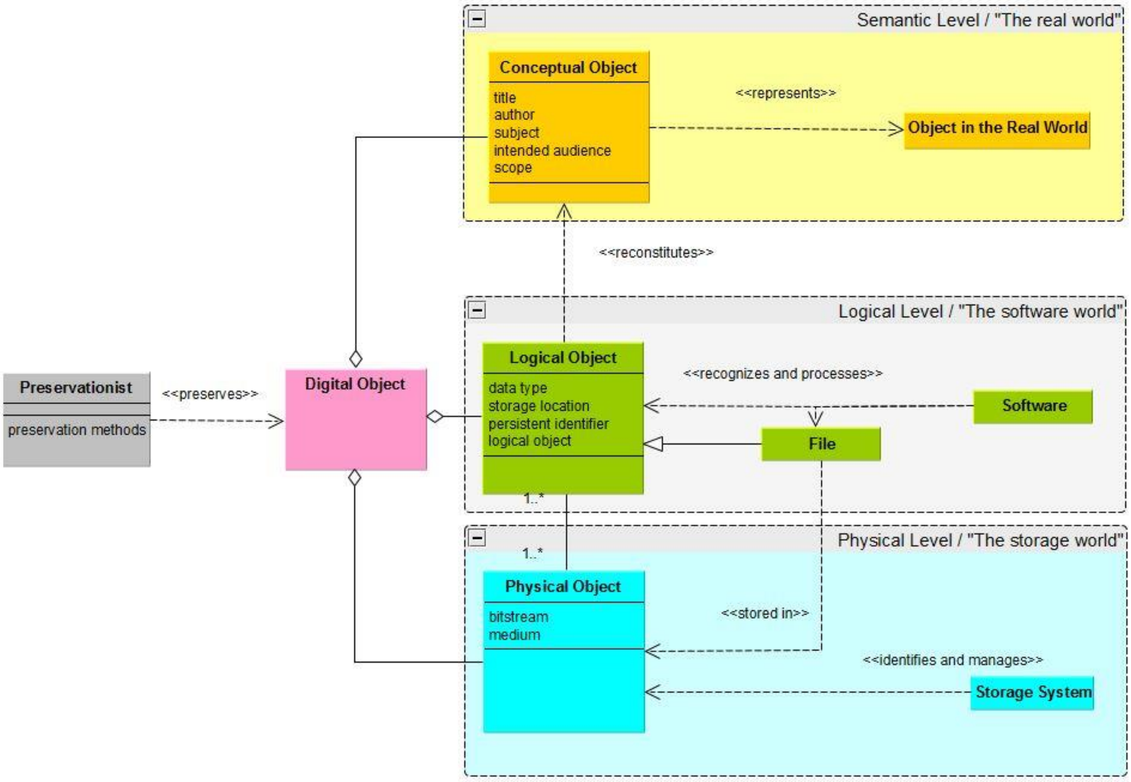


Figure 3: Object Levels of Preservation based on the works of Kenneth Thibodeau

Use Case - addressing the Complexity and Risks of Digital Objects via Object Levels of Preservation

In WissGrid, a German project (2009-2012) for providing data curation tools for a grid environment for research data, Thibodeau's model served as a basis for describing collective curation and preservation strategies by assigning responsibilities to actors of the RDM and the DP community and to the different technical infrastructures involved (Aschenbrenner, Ludwig et al., 2011; WissGrid, 2010). WissGrid redefined Digital Objects as *Digital Research Objects* and their bitstream, logical and semantic level as *Bitstream Preservation, Content Preservation* and *Data Curation,* which together are described as *curation levels* that aim towards maintaining/keeping the technical identity, and the technical and intellectual reuse of digital objects (see figure 3). This combined terminology can be seen as an approach to foster a shared understanding and awareness for different tasks (e.g. metadata extraction) and responsibilities when it comes to preserving research data. Three entities involved in the project were declared responsible for the three object levels, e.g. repositories of the D-Grid-Infrastructure were mainly responsible for bitstream preservation (see figure 5).

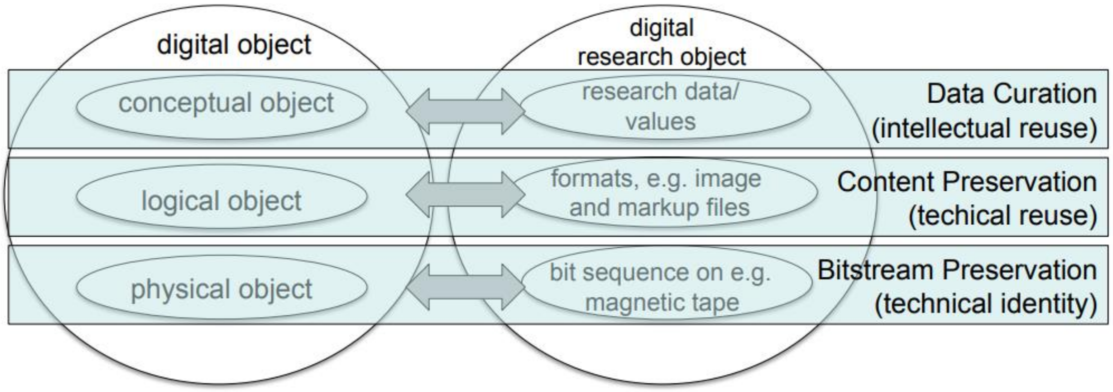

digital object
digital research object
conceptual object
research data/ values
Data Curation (intellectual reuse)
logical object
formats, e.g. image and markup files
Content Preservation (techical reuse)
physical object
bit sequence on e.g. magnetic tape
Bitstream Preservation (technical identity)

Figure 4: Curation Levels and Digital Objects in WissGrid (Ludwig, 2009

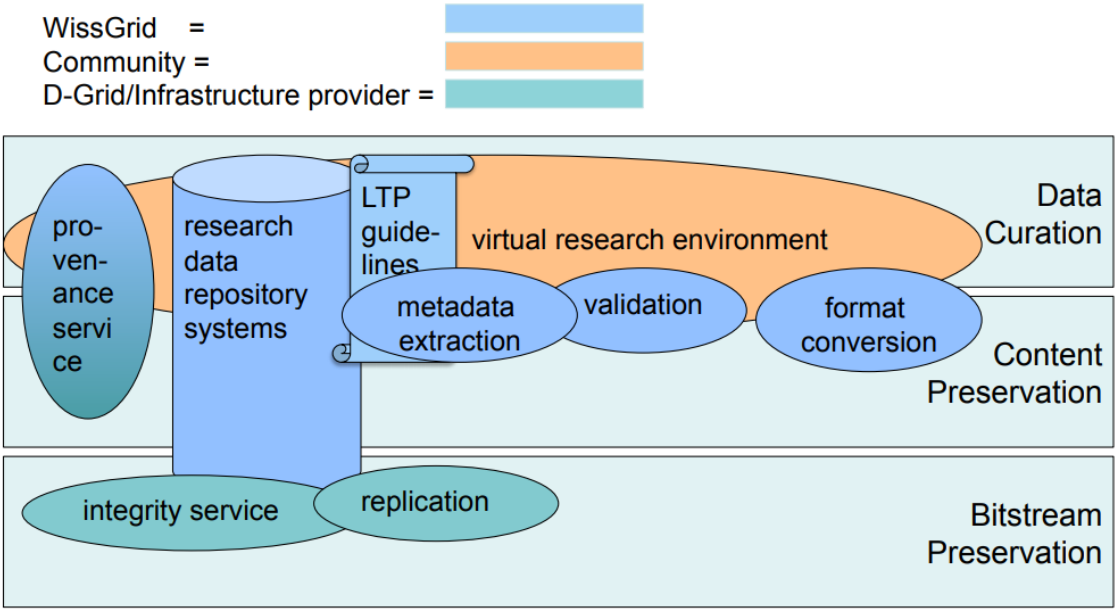


Figure 5: Responsibilities for Curation Levels in WissGrid (Ludwig, 2009)

Also, many institutions map their preservation processes against Thibodeau's object levels in practice. The three German National Subject Libraries TIB (2019), ZB MED (2019) and ZBW (2018) refer to them as one of their key principles in their preservation strategy. Within the digital archive of the three institutions, risks occurring at the bit-stream level are addressed via a solid managed storage infrastructure and regular integrity checking. Risks occurring at the logical level are addressed via file format characterization of preserved objects, the result of which is captured in preservation metadata, as well as preservation actions such as migration and emulation, which is based on the preservation metadata (see PREMIS). Lastly, risks associated with the semantic level are addressed via content and context description in descriptive metadata.

General Adoption and Critique

Nearly 18 years later, Thibodeau's model appears to still be of high relevance for the DP community (Chassanoff, Altmann; El Idrissi; Kylander et al.). The main value of Thibodeau's tripartite model has been seen in the differentiated look at Digital Objects and their complexity, resulting in the different layers which need to be considered in preservation actions (Kirschenbaum). Hence, the original Object Levels of Preservation physical, logical, and conceptual have been adopted in subsequent digital preservation models and partially renamed as B*it(stream)*, *Logical* and *Semantic Preservation* (DURAARK; Kylander et al.; Rauber). Use cases like WissGrid show that Thibodeau's Object Levels of Preservation are also of interest in RDM context.

Data Management Plans

While the last two core concepts presented models, we now move our focus to an actual process implemented in RDM practice. Data Management Plans (DMPs) can be seen as a route map towards reusable data which is effectively preserved and shared. While specific requirements and wordings vary, DMPs are designed to encourage content creators to reflect on the longer-term value of their **Data** and make appropriate plans to allow it to be preserved and reused. Indeed, most DMPs ask a greater number of questions on data sharing (how, when, with whom, any restrictions, etc.) than addressing simple data management concerns such as storage and backup. DMPs identify an envisioned location to preserve/share data and can help to identify obstacles that prevent preservation like identifying **Intellectual Property Rights**, **Ethics** and data protection issues, specialised or proprietary file formats. **Researchers** are expected to identify core **Metadata and Documentation** required to enable reuse so this can be created accordingly.

Key Terminology of the Model

While DMPs vary by institution and funder, there are common elements across all contexts (Williams et al. 2019). The DCC synthesised requirements and released a Checklist for a Data Management Plan (DCC, 2013). This proposed seven main topics:

- **Data** collection
- **Metadata and Documentation**
- **Ethics** and legal compliance
- **Storage and Backup**
- Selection and **Preservation**
- **Data Sharing**
- Responsibilities and Resources

A more recent publication from Science Europe (2019) on International Alignment of Research Data Management converges on very similar core requirements for DMPs, validating the key terminology. This proposes six rather than seven categories, merging **Data Sharing** and **Preservation**, and adding the concept of data quality to documentation.

The DCC also issued a set of common themes and associated guidance for DMPs. These were consulted on internationally, together with the University of California Curation Centre, and revised to a set of 14 themes (DCC & UC3, 2018). The themes are aligned to the topics noted above and specify more detailed concepts e.g. **Data Format**, data volume, **Persistent Identifiers** and **Data Repository**.

Use Case - DMPs as a bridge between creator and curator communities

Data Management Plans are typically created at a grant application stage or post-award. They are intended to address plans for the creation and management of data to ensure that it can be shared and preserved, as appropriate. DMPs are a useful talking point to bridge between content creation and curation communities. In many universities, tailored guidance and consultation services are offered to assist researchers to complete DMPs.

This helps to address challenging aspects of data security, ethics, licencing and preservation by raising awareness of relevant support services and ensuring best practice is followed.

There is a strong desire to increase connections between DMPs and repositories. Sharing information on expected data volumes so repositories know what is in the pipeline and can do capacity planning was a primary use case to emerge from consultative workshops at IDCC and Open Repositories (Simms et al., 2017 & Drafiova, 2019). Within NERC-funded research projects, the designated data repositories consult and co-creates the DMPs to ensure close alignment and better transition between the creation and curation roles.

There is an increasing trend towards publishing DMPs, either in journals such as Research Ideas and Outcomes, or by depositing in repositories. A survey of H2020 projects found that almost 50% were willing to openly publish their DMP and even more so if certain conditions such as confidentiality were met (Grootveld et al, 2018). DMPs provide context on the creators' intentions and choices and so offer useful insights to both curators and potential reusers. ZB MED - Information Centre for Life Sciences assists life scientists with creating and publishing DMPs within the projects RDMO4Life (2020) and EmiMin (2019). ZB MED is also planning to ingest DMPs as part of Submission Information Packages (see OAIS) in the digital preservation system Rosetta and thereby hold them available to users for the long term. This demonstrates the archival value DMPs themselves have since they document and contextualise the genesis of the preserved research data and can be seen as Representation

General Adoption and Critique

DMPs started gaining traction from c.2007 onwards with the increase in research funder policies encouraging or mandating plans (Jones, 2012). Initially they were predominantly conceived of as an administrative exercise or hurdle to obtaining funding. This interpretation was not helped by low levels of monitoring and follow-up if plans were not implemented. In recent years, the rhetoric of DMPs being 'living documents' which are continually updated throughout the course of research has increased, largely thanks to the European Commission policy (2016). There is also increased research activity on machine-actionable DMPs, seeking to facilitate information exchange across services in the research lifecycle e.g. sharing information from Research Information Management systems to prepopulate DMPs or extracting data volumes and preservation requirements to share with repositories (Simms et al., 2017).

Research funder data policy is increasingly placing an onus on research organisations to provide support services for data management and ensure intentions listed in DMPs are carried out. The UK's Engineering and Physical Sciences Research Council (EPSRC) policy released in 2011 was the first to acknowledge that research organisations are awarded funding and are ultimately responsible for ensuring data are managed and shared. Since then, the Arts and Humanities Research Council (AHRC) has released a number of data management points that institutions need to agree. These include confirming that the proposal has been written in line with the institution's data management policy and that the institution's data support (e.g. library services, IT department) have been consulted. Nordic funders such as the Swedish Research Council (2019) and Research Council Norway (2019) and Dutch funders such as Netherlands Organisation for Health Research and Development (2020) and the Netherlands Organisation for Scientific Research (2019) are expecting institutions to take responsibility for checking the DMP. The Health Research Board Ireland (2019) has provided training for institutional support staff as it hopes to receive validation that submitted DMPs have been reviewed and approved by them.

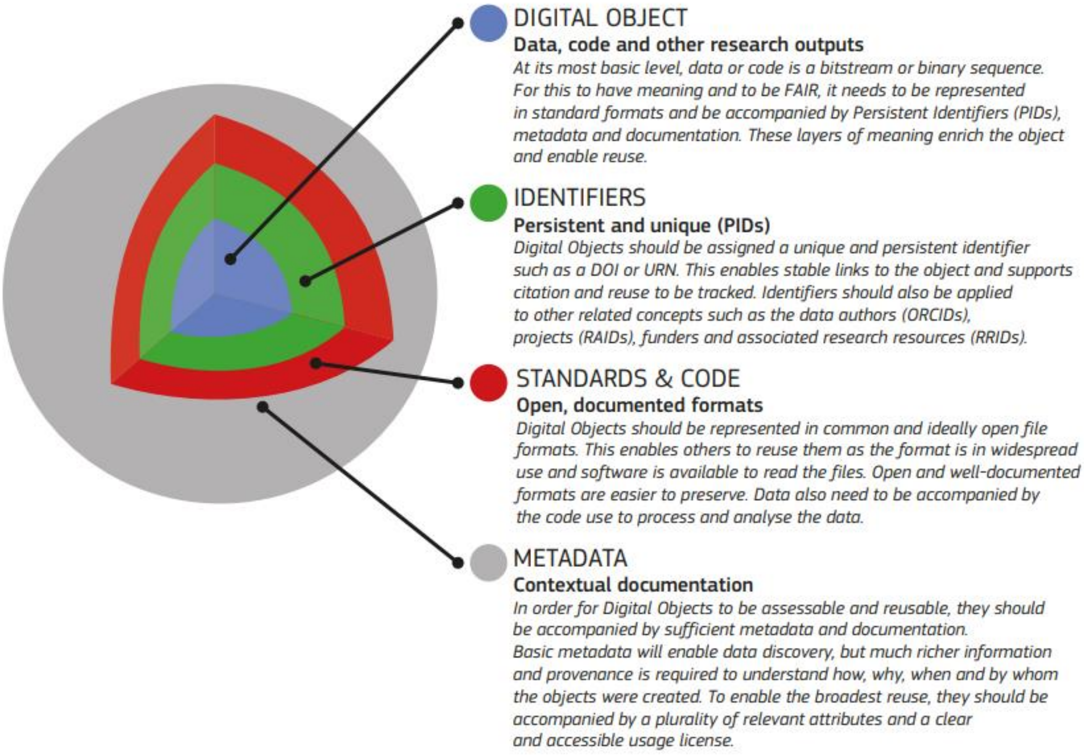
Since the topic of DP is often required and therefore usually addressed in DMPs (Williams et al., 2017), it is striking that DMPs have not had a significant impact on the DP community. Examples for DMP use cases in DP are rare. A theoretical example is given by Navale and McAuliffe who allocate DMPs in the Administration Functional Entity of OAIS (Navale & McAuliffe, 2018). Thus, DMPs seem to be of equal use for both RDM and DP, since they could help standardize the interaction between the Producer, the Archive and the Consumer side. Also, being part of a Submission Information Package (see OAIS), DMPs could add useful context to research data in a standardized manner and thereby help the Archive and Consumers understand the preserved research data.

FAIR

The concept of FAIR was conceived in the life sciences community at a Lorentz workshop in 2014. It is intended to encapsulate core principles for data and/or metadata, namely that they should be Findable, Accessible, Interoperable, Reusable (FORCE11). FAIR is being increasingly adopted in research funder policy and is an emphasis with the European Open Science Cloud initiative. As noted in the Turning FAIR into Reality Expert Group report (Hodson, Jones et al, 2018), the "FAIR principles focus on access to the data and do not explicitly address the long-term preservation needed to ensure that this access endures". The environment in which data are stewarded is fundamental so the report authors propose expanding the principles to address key concepts of DP.

Key Terminology of the Model

As shown on the FORCE11 website, FAIR comprises of four key concepts, namely making data **Findable**, **Accessible**, **Interoperable** and **Reusable**. Several terms predominate through the articulation of the steps to achieve these four concepts, namely metadata, vocabularies, standards, identifiers, protocols, licences and provenance, which are also addressed in DP models (see Figure 2). The core thrust of the FAIR principles is ensuring data are shared with associated metadata, licences and identifiers to enable reuse. Indeed, within the Turning FAIR into Reality Expert Group report, a model for FAIR **Digital Objects** is proposed and a basic minimum standard of FAIR is defined as **Discovery Metadata**, **Persistent Identifiers** and access to the data or metadata in **Standard Formats** under a clear **Usage Licence** (Hodson, Jones et al 2018). These are proposed as the key terms for the model.



DIGITAL OBJECT

Data, code and other research outputs

At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse.

IDENTIFIERS

Persistent and unique (PIDs)

Digital Objects should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE

Open, documented formats

Digital Objects should be represented in common and ideally open file formats. This enables others to reuse them as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

METADATA

Contextual documentation

In order for Digital Objects to be assessable and reusable, they should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the objects were created. To enable the broadest reuse, they should be accompanied by a plurality of relevant attributes and a clear and accessible usage license.

Figure 5: A model for FAIR Digital Objects proposed by the European Commission FAIR Data Expert Group (Hodson, Jones, 2018)

Use Case - Implementing FAIR throughout the health research lifecycle

The Health Research Board (HRB) Ireland is participating in a FAIR funder pilot which intends to make it easy for funders to require and support FAIR data (GO FAIR). The proposal intends to address the data management activities from proposal stage to sharing in seven steps:

1. Workshops are held to define metadata templates and FAIR metrics
2. Resulting metadata templates and FAIR metrics are made available via repositories
3. Funders select metadata templates to add as requirements to new calls
4. Researchers compose DMPs using tools with relevant metadata templates embedded
    a. Institutional data stewards receive alerts to approve DMPs
    b. Submitted DMPs are validated as approved by the institution
5. Funded researchers and data stewards execute the DMP
6. Data are deposited in repositories running automated FAIR metric evaluations
7. Trusted 3rd party evaluation services validate the FAIRness of the data and metadata
    a. Research funders receive automatic evaluation certificates

The HRB (2019) has supported research support staff from Irish institutions to attend data stewardship training and is piloting the DMPonline service for DMPs. GO FAIR is coordinating the study.

General Adoption and Critique

The FAIR data principles echo several earlier policies and emerging practice, encapsulating requirements in a memorable and desirable acronym. The main concepts don't propose anything new and align with practices amongst the better organised research communities. The astronomy community and linguists, for example, have defined a set of accepted data formats, metadata standards, sharing practices and infrastructure to facilitate sharing and reuse. Guidelines from groups such as the Australian National Data Service (now part of Australian Research Data Commons) on data transformation also align closely to the concepts put forward in FAIR.

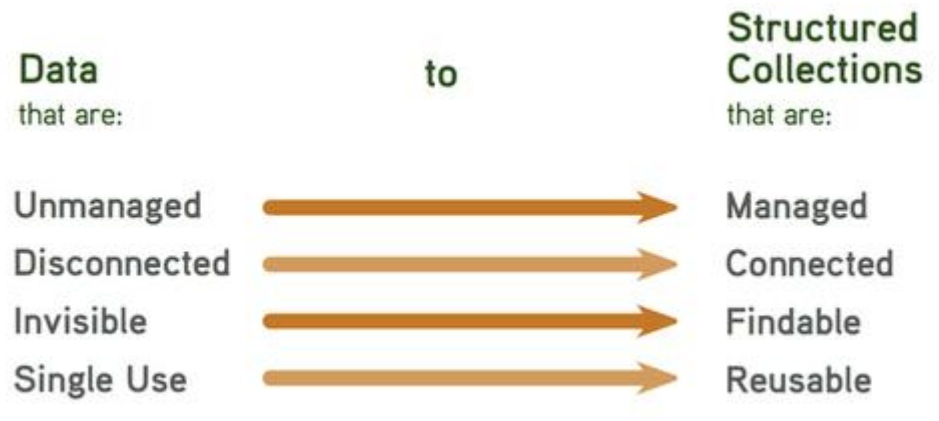


Figure 6 Data Transformations by the Australian National Data Service

The Royal Society report, Science as an Open Enterprise coined the term 'intelligent openness' to describe the preconditions for the effective communication of research data. It argued that being Open was not sufficient as data also need to be accessible, assessable, interoperable and usable (Royal Society, 2012). The 2013 G8 Science Ministers' Statement drew together properties mentioned in earlier policies, proposing that:

> Open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards (G8, 2013)

Two concepts which arise in these policies but are intentionally absent in FAIR are openness and data quality. FAIR does not mean open: data can be shared under access control or not at all - it is feasible to only provide metadata about a closed object. FAIR also says nothing on the quality of the object. The principles are only concerned with the technicalities of providing access, not offering any judgement on the quality of the object content or how this will be maintained. This is an area in which RDM and DP concepts can significantly strengthen FAIR. The reliability of research data is paramount and reproducibility scandals and fabricated data cases are rocking the research community. DP practice is also focused on ensuring the integrity of digital objects and not losing any significant properties (see OAIS) through preservation practice. A recent paper highlighted the benefits of combining concepts of Open data, FAIR and RDM (Higman et al, 2019) and Sierman has proposed opportunities to collaborate across DP and FAIR communities, suggesting that we could start with investigating whether FAIR Data Objects will lead to sustainable Archival Information Packages (Sierman, 2019) (see also OAIS).

Another core aspect currently missing in FAIR is trustworthiness - a gap which is currently addressed in the development of the TRUST principles: Transparency (T), responsibility (R), user community (U), sustainability (S) and technology (T). These are proposed as core principles required to keep data FAIR over time via a network of trustworthy digital repositories (Dawei et al, 2019). Like FAIR, the TRUST model is intended as a system of generic high-level metrics, which need to be mapped to other models such as OAIS or certification processes like CoreTrustSeal. Sierman points out that with FAIR being an RDM based model, a clear hand-over to a digital archive for long-term sustainability is currently missing and proposes "phase 1" as the research life cycle and "phase 2" for DP. Following this stream of thought research data repositories become Producers themselves, depositing data to DP, making the importance of considering the long-term impact of FAIR decisions from the get-go eminent.

One curatorial motivation for FAIR could be to mitigate issues that make acquisition and ingest of data into an archive difficult and expensive, particularly when data is offered through self-deposit platforms. These are problems like insufficient documentation to describe the methodology of data collection, of data processing or cleansing which impacts the quality of metadata about the data set. The same holds true for actions the researcher might undertake without keeping potential reuse scenarios in mind (the "What would someone need to understand my data" issue) - e.g., inconsistent or meaningless file names, proprietary formats or corrupt files which can pose a real challenge to accessibility and preservation. Introducing DP processes such as file format identification and validation upstream at the point of creation could significantly ease the hand-over between RDM and digital preservation repositories.

Reference Model for an Open Archival Information System (OAIS)

The Reference Model for an Open Archival Information System (OAIS, ISO 14721) is the standard for the DP community, describing the core entities and functions an Archival Information Systems needs to include (CCSDS. 2012). While the first version was released in 2002, the current version dates to 2012, a new revision is to be released in 2021, following a public consultation period via a Consultative Committee for Space Data Systems (CCSDS) red book release in 2020 (Kearney, 2019). As a reference model the OAIS is not a blueprint for a system but rather a conceptual description of functional components, how they relate to each other and how they process information objects with the objective of long-term availability.

Key Terminology of the Model

An Open Archival Information System is defined as "an **Archive**, consisting of an organization, which may be part of a larger organization, of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community" (ISO 14721). The **Designated Community** is an identified group of potential **Consumers** who should be able to understand the information over time.

The actual target of preservation is the **Content Information (CI)** which consists of a **Digital Object** and accompanying **Representation Information**, i.e. information which describes the structural or semantic context of the digital object.
In order to preserve and make this information available, the Archive interfaces with three external roles: **Producer** and **Consumer**, who can be people or systems and provide /

consume information to / from the system, and **Management** who provides the overarching managerial / policy context the OAIS exists in. An example for Management can be a university who maintains an OAIS but naturally also has other functions.
The day-to-day management of the OAIS is described one of the six functional entities (i.e., **Administration Functional Entity**). Each functional entity describes services and functions the Archive should provide to fulfil the goal of preservation. The **Ingest Functional Entity** describes accepting data from the Producer and preparing it for storage and management within the Archive. From there data is passed to the **Archival Storage Functional Entity** from where it is made findable and accessible by the **Access Functional Entity**. The population and maintenance of descriptive and administrative data required to manage objects in the Archive are the responsibility of the **Data Management Function**. Lastly, the **Preservation Planning Functional Entity** describes the services and functions to ensure that the data is available and interpretable over the long term in the face of ongoing technological changes. This includes tasks like regular evaluation of the Archive's contents, risk assessment, monitoring of the changing technological environments and Designated Community expectations, risk assessment and migration or emulation plans.

As the CI moves through the functional entities additional information may be added or only a sub-view may be generated, e.g. when giving access to only specific aspects of a dataset based on legal access restrictions. In order to meaningfully describe how these different logical container versions of Content Information move through the system, the OAIS uses a logical **Information Package** concept. An information package consists of the CI as well as additional **Preservation Description Information (PDI)** such as provenance, access rights or fixity information. Three versions of information packages exist: the **Submission Information Package (SIP)** is supplied to the Archive by a Producer and moves through Ingest to become an **Archival Information Package (AIP)**, which, in return, is stored in the Archival Storage entity with the goal to be - in its entirety, as a subset, or as a combination with other AIPs - delivered to Consumer as a **Dissemination Information Package (DIP)**.

Over the course of time changes to digital objects or the infrastructure they are typically consumed with may change. Due to this it is important to define **Transformational Information Properties** (sometimes synonymously described as **Significant Properties,** see also PREMIS) which describe aspects of the Content Information that need to be preserved. Such properties can relate to an object's appearance, such as a colour space as well as behaviour, such as the possibility to edit a document.

Use Case - OAIS in the CoreTrustSeal Certification Process

As the standard reference model for DP processes, OAIS forms a natural basis for trustworthy archive certification processes. Currently three different certification processes exist - ISO 16363 "Audit and certification of trustworthy digital repositories", which is authored by the same working group which maintains OAIS, the nestor Seal which is based on DIN 31644 and the CoreTrustSeal (CTS) certification process, which has a particularly strong foothold in the RDM community. CTS criteria heavily rely on OAIS key concepts. CTS requires its applicants to describe processes derived from functional entities such as archival storage (R9), preservation planning (R10) and data management (R7). While the aforementioned key terminology makes indirect references to the OAIS, the standard is explicitly mentioned in combination with technological infrastructure. In particular in the storage requirement (R9) the guidance states that "Repositories that perform digital preservation must offer 'archival storage' in OAIS terms" (CTS, 2019).

An OAIS concept used throughout all CTS is the Designated Community: archives undergoing certification are asked to supply a description of their Designated Community in R0. Subsequent requirements such as Appraisal (R8) require implemented processes to match Designated Community requirements.

General Adoption and Critique

In DP, OAIS is omnipresent. One of the standard's biggest merits is that it has given the community a vocabulary to describe implementations and processes. However, this does not make OAIS easy to digest and critique exists. In particular the Designated Community concept is one which is controversially discussed. Several aspects make the concept highly speculative in nature - first off, the Designated Community are not actual users who access the data today but are potential consumers who should understand the data that is being preserved. In addition, assumptions about this community's specific knowledge base and scope needs to be made. It seems logical that such definitions are a necessary basis for preservation action boundaries within an Archive. In practical use, however, this is especially a problem for large archives who consider the general public their main Designated Community, for whom a knowledge base is almost impossible to define (Lindlar and Rudnik, 2019).

While OAIS is deeply embedded in the DP community, the same cannot be said for the RDM community. An analysis of 40 CTS self-assessment reports conducted in 2019 has put forth gaps in the application of OAIS terminology such as Archive and Designated Community (Lindlar and Rudnik 2019). The term Designated Community is not widely used within the RDM community, instead, "the community" RDM support staff work with is a wide ranging group (Wilkinson et al. 2016). But for this topic, we can implicitly narrow "the community" to the discipline in which researchers work and the rules, norms, and expectations governing the collection, use, and reuse of data in the context of their research field.

PREMIS

PREMIS (PREservation Metadata Implementation Strategies) is a de-facto standard for metadata needed to understand data across its lifecycle within an archive (PREMIS Editorial Committee, 2015). While the development of the data dictionary dates back to 2003, the current version 3.0 of the community maintained[2] standard was released in 2015. In addition to the data dictionary, a PREMIS OWL Ontology, an XSD schema as well as information about different implementations are made available through the PREMIS website[3].

PREMIS is considered a subset of all Preservation Metadata, namely descriptive metadata; detailed info on agents, rights, media or hardware; format-specific technical metadata and repository business rules (Caplan, 2009).
It is important to note that PREMIS does not prescribe a specific implementation - PREMIS conformance can be achieved via the ability to map to the data dictionary entries[4].

Key Terminology of the Model

[2] https://www.loc.gov/standards/premis/premis-editorial-committee.html

[3] https://www.loc.gov/standards/premis/v3/index.html

[4] http://www.loc.gov/standards/premis/premis-conformance-20150429.pdf

The PREMIS data model defines 5 core semantic entities which play a role in understanding data: **(Digital) Object**, **Event**, **Agent**, **Rights Statement** and **Environment**[5].

In addition, the Object entity is broken down into four subsequent subcategories: **Intellectual Entity**, **Representation**, **File** and **Bitstream**. While the Intellectual Entity describes the "distinct intellectual or artistic creation that is considered relevant to a **designated community** in the context of digital preservation" (PREMIS 2015), such as a book or a research **data set**, a representation is a complete rendition of that intellectual entity, e.g. a PDF version of the book / research data set or an XML version of the book / research data set. As such, each representation contains one or more files which may be broken down further into bitstream, e.g. in the case of a PCM audio stream within an audio-visual container.

Core knowledge about both the digital object as well as about any action performed on that object can be mapped to these semantic entities and is described within the data dictionary in 88 semantic units. For example, the Object Entity may include information about how the digital object was created (creatingApplication semantic container and subunits) or what characteristics are deemed to be **Significant Properties** (see also OAIS), i.e., characteristics which need to be preserved over the course of preservation action such as migration (signifcantProperties semantic container and subunits). The Agents Entity may include information about a specific tool (agentName, agentType, agentVersion) which was used to migrate the object into a new format while the Event entity, in return, protocols information about the type of action (eventType) and the time it was performed (eventDateTime) as well as the outcome (eventOucomeInformation). while the environments entity includes information about the hardware and software (environmentName, environmentVersion, environmentFunction) required to correctly render the digital object.

Use Case - PREMIS in a large-scale digital preservation system

Most digital preservation systems implement PREMIS - this includes large commercial products like Preservica, Archivematica or Rosetta (PREMIS Implementation Registry, 2018). The standard does not prescribe a specific implementation, therefore allowing for various different degrees and forms of how it is used in practice.

Within Rosetta, PREMIS is implemented in the ExLibris-defined DNX metadata schema. In accordance with the PREMIS conformance statement (PREMIS Editorial Committee, 2015b) a mapping between Rosetta DNX and the PREMIS Data Dictionary is supplied in the publicly available Rosetta AIP Data Model Guide (Ex Libris, 2019).

Information gathered as part of a PREMIS implementation with a system can be leveraged in many digital preservation processes downstream. One example is the PREMIS semantic unit container inhibitors: It might become necessary for a repository to store a password protected object. While some archives choose to reject such objects, others may not be at liberty to do so - or may not want to completely reject the object as the information contained within is indeed highly relevant to the collection the archive is preserving. In such a case one would want to capture information that the object is password protected and - if available - even store the password as well. PREMIS allows to

[5] Strictly speaking, Environments are described largely reusing the Object entity. However, due to special environment container units within the Object entity, as well the fact that they can be linked to other Objects and Agents, they are considered a semantic entity in their own right.

capture this information in the inhibitors container, where the inhibitorType unit indicates that an object is password-protected, the inhibtorTarget unit may indicate what is restricted, e.g. print only, and the inhibitorKey can even store the password itself. The implementation of the container in a schematized manner in Rosetta allows the users to query the system for all objects for which a specific inhibitor exists and e.g. describe them in an according preservation plan (for further examples, see: Lindlar, 2018).

General Adoption and Critique

The PREMIS Implementation Registry currently lists 50 entries by institutions who briefly describe their usage of PREMIS. Entries range from large National Libraries and Archives over those made by software vendors or open-source developers to research data repositories. The list clearly shows that there is a wide adoption of the data dictionary and PREMIS can be implemented by archives of different size, content and domains. The role of PREMIS within the DP community is further underlined by the OAIS pointing to the PREMIS data dictionary as a suitable standard to follow in the submission of digital metadata, about digital or physical data sources, to the Archive (CCSDS, 2012).

Terminology Mapping and Discussion

In a second step, we explore the terminology extracted in the analysis and identify where the analysed models use different terms for the same concept or use the same terms for different concepts, and where concepts from one community can be beneficially applied to another.

The terminology and mapping is documented in the table presented in Annex A, and has been clustered in four groups which we could identify while discussing the mapping. The discussion presented here follows this clustering.

What is being managed or preserved?

The first group of terminology represents the “what”, i.e., the actual target that is being managed or preserved in RDM and DP processes. While some concepts such as OAIS can be applied to analogue targets of preservation, we limited the terminology within the scope of this paper strictly to digital. In the majority of concepts across both domains the main target is referred to as the **Digital Object** - DMPs define this broader as **Data**, while OAIS defines this more specific as an **Information Package** (SIP, AIP, DIP).
The DP models OAIS, PREMIS and Preservation levels each present their own granular understanding of the **Digital Object**, not contradicting but complementing each other. DP models take the genesis of the **Digital Object** and its accompanying metadata as it moves through different processes into account, i.e. through different **Information Packages** in the OAIS.

Thibodeau’s superordinate concept of **Digital Object** is comparable to the semantic entity **Object** in PREMIS and to the different types of **Information Packages** in OAIS. Each DP model acknowledges that any **Digital Object** needs to be broken down into different sub-objects which carry their own properties that are relevant for preservation. The **Physical Object** according to Thibodeau is similar to the **Digital Object** in OAIS and to **Bitstream** in PREMIS, whereas the **Logical Object** can be mapped to **Representation** and **File** in PREMIS and to **Content Information** in OAIS. Thibodeau’s **Conceptual Object** corresponds with **Intellectual Entity** in PREMIS, but has no equivalent in OAIS.

While RDM considers a **Digital Object** in its entirety, not using separate terminology for the physical aspect or the conceptual content, a mapping is often understood on the level of the **Logical Object**, meaning the functional layer of a **Digital Object**, i.e., its file format or encoding. Based on the vocabulary, this leads to a mapping of the RDM terminology **Data Formats** (DMP) and **Standard Formats** (FAIR) to the **Logical Object**. However, discussion amongst the authors highlighted that in RDM these concepts are mainly understood as the need for sustainable file formats to support long-term accessibility of the **Digital Object**. Here, DP models go one step further, considering the functional properties of the logical layer, or the file format itself.

Contextual information about the target of Research Data Management or Preservation

Information about the target, or the **Digital Object**, is referred to as **(discovery) metadata** (DMP), **metadata and documentation** (FAIR) or **Representation Information** (DCC, OAIS) in the majority of the models. As the Object Levels only describe the **Digital Object** itself, metadata can be mapped to the **Properties of all Object Classes** while PREMIS is a metadata standard in itself. Both, RDM and DP models, recognize that different classes of information need to be captured about the **Digital Object**. Information about rights and identifiers are two categories that are found across both domains.

A unique information category of the DP domain results in the aforementioned consideration of a file format as an information carrying layer with relevance to preservation. The consideration of functional properties of the logical layer, but also on other layers, is captured in something currently unique to the DP domain: the concept of **Significant Properties** (PREMIS) or **Transformational Information Properties** (OAIS). These properties define unique features of a **Digital Object**, meaning the object on all layers in the DP sense, which need to be maintained across preservation action such as migration. An example is a word document which is migrated to a PDF file. If a **Significant Property** were that the object must remain editable and track changes should be kept in place, a migration to PDF would not preserve these behavioural features of the **Digital Object**.

Discussion between the authors identified that such a concept is currently missing in RDM, even though the identification of **Significant Properties** by the object's creator would be of great help in the preservation process. Researchers should liaise with curators about how their data is likely to be used to ensure preservation processes don't lose key characteristics and functionality.

Environment and Actors

With the exception of the Object Levels of Preservation model, which mainly deals with the Digital Object, all analysed models recognize the Digital Object's environment and its corresponding actors, however, to a different extent. In PREMIS all actors are defined as **Agents**, which, in return, can be persons, organizations or software. The other models differentiate by role or process. The creator of the Digital Object is described as a **Producer** (OAIS) or a **Researcher** (DMP), the entity accessing the object is described as a **Consumer** (OAIS) or a **Data User** (FAIR). The DCC Lifecycle Model is agent agnostic, inferring to them via the respective actions **Create / Receive** and **Access / Reuse**, which are partially mapped in the process & functions category.

The Environment and Actors category underlines the creator-orientation of RDM and consumer-orientation of DP in two terminologies:

Ethics is a strong concept in DMPs, describing the circumstances under which data was gathered and is to be treated in the future. It is a clear statement towards a socio-environmental awareness that is captured and defined at the point of creation. It is also a terminology which does not come up on the other models. While we found it best mapped to **Appraise & Select** (DCC Lifecycle) and **Management** (OAIS), it proposes an in-depth look at the social, rather than the technical environment which the object was created in - a thought currently lacking in DP models.

While the **Designated Community** (OAIS, PREMIS) is terminology which only exists in DP models and considers the (future) user whom the archive is preserving the object for, the DCC Lifecycle Model indirectly references it via **Community Watch**. The underlying idea is that requirements of the user community are monitored and cross-checked against processes implemented in the archive to ensure that Digital Objects fulfil the actors' needs. DMPs sometimes ask researchers to propose likely reuse scenarios. Such a definition of the audience for whom the Digital Object was created / deemed relevant as part of a DMP definition, would allow for a better tailoring of preservation action downstream.

Processes / Functions

In PREMIS processes are defined as **Events** which are performed on the Digital Object, the outcome of which is documented. Two concepts with strong process and function definitions are the OAIS with its definition of **Functional Entities** and the process-centred DCC Lifecycle Model.

FAIR rather describes the "what" than the "how" - indicating that Digital Objects need to be **Accessible**, but not really indicating how that should be achieved. In particular, a mapping to preservation and storage is missing in FAIR. DMP, on the other hand, does mention **Storage and Backup** as well as **Preservation**. While we did map these to the OAIS **Archival Storage Functional Entity** and the **Preservation Planning Functional Entity**, the analysis has shown that **Preservation** and **Storage** in DMPs is often understood as being solved by hand-over to a repository without taking into consideration whether said repository actually includes preservation.

Another case where a misalignment of understanding may exist is in the mapping of DMP's **Data sharing** to **Access**. Whereas RDM seems to understand **Data Sharing** as being synonymous to **Access** and also to openness, that reading is a one way street. **Access** in general can be conditional - it can be internal, retention period or trigger-event based or otherwise limited to a specific audience.

Conclusion

We embarked on this journey by asking the question of whether the DP and RDM communities, who converge at conferences like IDCC or iPRES, in fact do speak the same language, thus enabling the use of synergies via collaboration, or whether they are rather coexisting. Working on this paper, the authors - who are representatives of the RDM and the DP communities themselves - analysed several key concepts and extracted - sometimes RDM or DP specific - terminology that are of use for both communities. While some terminology may be uncommon within the DP or the RDM community, they often describe the same or similar ideas and thereby complement each other in many cases.

What unites both the RDM and the DP community and what is apparent in all models, is its main subject of interest: the digital object and its corresponding metadata. Within the models, different terms and levels of specificity are used, but whether researchers are creating data or curators are managing and preserving it, an awareness of the object and associated metadata to provide meaning is paramount. Our analysis shows that DP models have a more granular understanding of a Digital Object than RDM models. This may, in some cases, lead to misunderstandings, e.g. using different terms for the same concept of Digital Object or vice versa (e.g. different meanings of Digital Object in OAIS and the DCC Lifecycle Model). Although RDM models like the DCC Lifecycle Model also put the Digital Object in the centre, they do not provide a similar differentiated look at its specific characteristics. Our mapping can be seen as a first step to transfer the granular understanding of Digital Object into the RDM community.

Our recommendation for collective curation and preservation strategies is to implement this object-cantered view. Two examples of where this could be useful in particular are the awareness of file formats as an information carrying choice and the definition of Significant Properties as well as a clearer modelling of how different Files within a data collection tie together into Representations, i.e., the same content presented in a different form, especially if different representations are present within a Digital Object / data collection, e.g., a digitized artifact in two different point-cloud file formats.

In mapping DP and RDM concepts and their respective terminology to each other, exploring where they overlap and where their definitions may lead to different interpretations, a first step towards a shared understanding has been made. Such a shared understanding shall pave the way towards collective curation and preservation strategies. The analysis has shown that there is indeed a significant number of overlaps, but also areas in which the communities can learn from one another. These include:

- Applying the DMP concept in digital preservation to encourage early engagement with content creators and consider what needs to be preserved
- Apply the Designated Community concept in RDM. It would be useful to liaise with researchers to understand who is most likely to use the data and how, using this information to inform choices of file formats, standards and preservation approaches.
- Apply digital preservation concepts to FAIR to ensure Digital Objects remain usable over time and are effectively preserved.

In the course of writing this paper we have identified several overlaps in the models we apply and complementary terminology and concepts. Much can be learned by closer connection of our communities. Ultimately, we all have the same end goal – there is more that unites than divides us.

Appendix A

	DMPs	FAIR	DCC Lifecycle Model	OAIS	PREMIS	Object Levels of Preservation
What is being managed / preserved?	Data	Digital Objects	Digital Objects	Information Package (SIP, AIP, DIP)	(Digital) Object	Digital Object
				Digital Object	Bitstream	Physical Object
	Data Format	Standard Formats		Content Information	Representation File	Logical Object
					Intellectual Entity	Conceptual Object
Contextual information about target	Metadata and Documentation	Metadata Interoperable	Representation Information	Representation Information Preservation Description Information	Semantic Units	Properties of Object Classes
				Transformational Information Property / Significant Property	Significant Property	
	Persistent Identifiers	Findable / Persistent Identifiers		Data Management Functional Entity		
	Intellectual Property Rights	Usage Licence			Rights Statement	
Environment / Actors	Data Repository				Environment	
				Archive		
			Community Watch	Designated Community	Designated Community	
	Researcher		Create/receive	Producer	Agent	
		Reusable / Data User		Consumer	Agent	
			Appraise & select	Management	Agent	
Processes / Functions				Administration Functional Entity	Event	
			Ingest	Ingest Functional Entity	Event	
	Storage and Backup		Store	Archival Storage Functional Entity	Event	
	Preservation		Preservation Planning	Preservation Planning Functional Entity	Event	
	Data Sharing	Accessible	Access / Use / Reuse	Access Functional Entity	Event	
			Preservation action Transform		Event	

References

[Website] Arts and Humanities Research Council (n.d) Text for Funding guide https://ahrc.ukri.org/documents/guides/data-management-plan/

[Article] Aschenbrenner, A., Ludwig, J. et al. (2011). *Diversity and Interoperability of Repositories in a Grid Curation Environment,* in Journal of Digital Information 12(2). *https://journals.tdl.org/jodi/index.php/jodi/article/view/1896/1770*

[Report] Caplan, P. (2009). *Understanding PREMIS*. Library of Congress Network Development and MARC Standards Office. https://www.loc.gov/standards/premis/understanding-premis.pdf, p. 5

[book] CCSDS (2012). *Reference Model for an Open Archival Information System (OAIS)* - Magenta Book, CCSDS.[Online]. Available: http://public.ccsds.org/publications/archive/650x0m2.pdf.

[Article] Chassanoff, A., Altman, M. (2019). *Curation as "Interoperability With the Future": Preserving Scholarly Research Software in Academic Libraries*, in: Journal of the Association for Information Science and Technology, May 2019, https://doi.org/10.1002/asi.24244

[Report] CTS (2019). Core Trustworthy Data Repositories Requirements. https://www.coretrustseal.org/wp-content/uploads/2019/11/2019-10-CoreTrustSeal-Extended-Guidance-v2_0.pdf

[Report] Dawei, L., Crabtree, J. et al (2019). *The TRUST Principles for Digital Repositories - A White Paper. Version 0.03 (draft).* https://bitly.com/trustprinciples

[Report] DCC (2013) Checklist for a Data Management Plan. v.4.0. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/data-management-plans

[Website] DCC & UC3. (2018) Themes for Data Management Planning https://github.com/DMPRoadmap/roadmap/wiki/Themes

[Website] Deutsche Forschungsgemeinschaft (2018) Handling of Research Data https://www.dfg.de/en/research_funding/proposal_review_decision/applicants/research_data/index.html

[Website] Drafiova, M. (2019) Closing the gap – connection points between DMPs and repositories. DCC blog. http://www.dcc.ac.uk/blog/closing-gap-%E2%80%93-connection-points-between-dmps-and-repositories

[Report] DURAARK (2014). *D6.6.1 Current state of 3D object digital preservation and gap-analysis report*. https://zenodo.org/record/1115504

[Proceeding] El Idrissi, B. (2019). *Long-Term Digital Preservation: A Preliminary Study on Software and Format Obsolescence.* Proceedings of the ArabWIC 6th Annual International Conference Research Track, Article No. 13, https://dl.acm.org/citation.cfm?doid=3333165.3333178

[Website] EmiMin (2019) Verbundvorhaben Emissionsminderung Nutztierhaltung (EmiMin), https://www.ktbl.de/themen/emimin/

[Website] EPSRC 2015 EPSRC policy framework on research data Expectations https://epsrc.ukri.org/about/standards/researchdata/expectations

[Report] European Commission (2016) Guidelines on FAIR Data Management in Horizon 2020 https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[Report] Ex Libris Limited (2019). Rosetta AIP Data Model. https://knowledge.exlibrisgroup.com/@api/deki/files/78233/Rosetta_AIP_Data_Model.pdf?revision=1

[Website] GO FAIR (no date). FAIR funding cycle https://www.go-fair.org/today/fair-funder

[Website] FORCE11. (no date) FAIR data principles https://www.force11.org/group/fairgroup/fairprinciples

[Website] G8 Science Ministers Statement, 13 June 2013 https://www.gov.uk/government/news/g8-science-ministers-statement

[Website] Government of Canada (2019) Frequently Asked Questions Tri-Agency Research Data Management Policy - Science.gc.ca http://www.science.gc.ca/eic/site/063.nsf/eng/h_97609.html

[Report] Grootveld, M. Leenarts, E. Jones, S. Hermans, E & Fankhauser E. (2018). OpenAIRE and FAIR Data Expert Group survey about Horizon 2020 template for Data Management Plans http://doi.org/10.5281/zenodo.1120245

[Website] Health Research Board of Ireland (2019) Policy on Management and Sharing of Research Data https://www.hrb.ie/funding/funding-schemes/before-you-apply/all-grant-policies/hrb-policy-on-management-and-sharing-of-research-data/

[Article] Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation.* 3(1), p. 134-140. https://doi.org/10.2218/ijdc.v2i2.30

[Article] Higman, R. Bangert, D. & Jones, S. (2019). Three camps, one destination: the intersections of research data management, FAIR and Open. Insights, 32(1), p.18. http://doi.org/10.1629/uksg.468

[Report] Hodson, S., Jones S. et al. (2018). *Turning FAIR into Reality: Final Report and Action Plan from the European Commission Expert Group on FAIR data*. European Commission. https://doi.org/10.2777/1524

[Article] Jones, S. (2012) Developments in Research Funder Data Policy. International Journal of Digital Curation, 7(1). https://doi:10.2218/ijdc.v7i1.219

[dataset] Kearny, M. (2019). *DAI WG Integrated Schedule v20181019*. https://cwe.ccsds.org/moims/_layouts/15/WopiFrame.aspx?sourcedoc=/moims/docs/MOIMS-DAI/DAI%20Schedule%20Overview/DAI%20WG%20Integrated%20Schedule%20v20181019.xlsx&action=default

[book] Kirschenbaum, M. (2008). Mechanisms: New Media and the Forensic Imagination, Washington, DC. Available: https://books.google.de/books?id=CT0oPmcrciAC&pg=PA4&lpg=PA4&dq=thibodeau+logical+conceptual&source=bl&ots=uQi7BGAKuI&sig=ACfU3U3mHLk7kmkryPt8vb5XyK2eay1Qig&hl=de&sa=X&ved=2ahUKEwjX4L-A-6rmAhVCcZoKHX3KBmAQ6AEwAXoECAkQAQ#v=onepage&q=thibodeau%20logical%20conceptual&f=false

[Proceeding] Kylander, J., Helin, H. et al (2019). *Together Forever, or How We Created a Common and Collaborative Digital Preservation Service*. Proceedings of the 16th International Conference on Digital Preservation, iPRES 2019, Amsterdam, September 16-20, 2019, https://ipres2019.org/static/proceedings/iPRES2019.pdf , p. 290-296.

[Slides] Lindlar, M. (2018). *PREMIS in Rosetta - A Play in 3 Acts.* https://doi.org/10.5281/zenodo.3626828

[Article] Lindlar, M., Rudnik, P. (2019). *Eye on CoreTrustSeal - Recommendations for Criterion R0 from Digital Preservation and Research Data Management Perspectives*. In: Proceedings of the 16th International Conference on Digital Preservation. iPRES2019.

[Slides] Ludwig, J. (2009). *Long-term Preservation of Digital Research Data* http://www.desy.de/dvsem/WS0910/ludwig_talk.pdf

[book] Müller, U., Scholze, F. et al. *DINI Certificate for Open Access Repositories and Publication Services 2016*. Editor: DINI Deutsche Initiative für Netzwerkinformation. http://dx.doi.org/10.18452/18178

[Article] Navale, V., McAuliffe, M. (2018). *Long-term preservation of biomedical research data,* https://doi.org/10.12688/f1000research.16015.1

[Website] Netherlands Organization for Health Research and Development (2020) Research Data Management in Your Project https://www.zonmw.nl/en/research-and-results/fair-data-and-data-management/data-management-in-your-project/

[Website] Netherlands Organisation for Scientific Research (2019) “NWO to update its data management protocol in January 2020” https://www.nwo.nl/en/news-and-events/news/2019/12/nwo-to-update-its-data-management-protocol-in-january-2020.html

[Website] PREMIS Implementation Registry (2018), https://www.loc.gov/standards/premis/registry/index.php

[book] PREMIS Editorial Committee (2015). *PREMIS Data Dictionary for Preservation Metadata*. Version 3.0. [Online] Available: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

[Report] PREMIS Editorial Committee (2015b). Conformant Implementation of the PREMIS Data Dictionary.

[Slides] Rauber, A. (2013). *Digital Preservation Introduction, http://www.ifs.tuwien.ac.at/~becker/slides/2013-Rauber-02_dp-intro.pdf*.

[Website] RDMO4Life (2020) (hosted by ZB MED - Information Centre for Life Sciences), https://rdmo.publisso.de

[Report] Royal Society (2012), Science as an Open Enterprise https://royalsociety.org/policy/projects/science-public-enterprise/Report

[Website] Research Council of Norway (2019) Data Management Plan https://innsida.ntnu.no/wiki/-/wiki/English/Data+management+plan

[Report] Science Europe. (2018) *Practical Guide to the International Alignment of Research Data Management.* https://www.scienceeurope.org/media/jezkhnoo/se_rdm_practical_guide_final.pdf

[Article] Simms, S., Jones, S., Mietchen, D. & Miksa, T. (2017) *Machine-actionable data management plans (maDMPs)* Research Ideas and Outcomes 3: e13086. https://doi.org/10.3897/rio.3.e13086

[Website] Springer Nature (2019) Research Data Policies FAQs https://www.springernature.com/gp/authors/research-data-policy/data-policy-faqs

[Website] Sierman, B. (2019) Do FAIR data ever become heritage? https://digitalpreservation.nl/seeds/do-fair-data-ever-become-heritage/

[Article] Sveinsdottir, T. Jones, S. Hwang, H, Kim, J & Rhee, H. (forthcoming) Implementing a Content Curation Lifecycle Model at KISTI in International Journal of Digital Curation.

[Website] Swedish Research Council (2019) Producing a data management plan https://www.vr.se/english/applying-for-funding/requirements-terms-and-conditions/producing-a-data-management-plan.html

[Report] Thibodeau, K. (2002). *The State of Digital Preservation: An International Perspective, chapter Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years*, Report Number 31, CLIR

[Website] TIB (2019) *Preservation Policy of the Technische Informationsbibliothek (TIB) – German National Library of Science and Technology,* https://www.tib.eu/en/service/tib-preservation-policy/

[Article] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

[Article] Williams, M., Bagwell, J., Zozus, M. (2017). *Data Management Plans, The Missing Perspective.* Journal of Biomedical Informatics, 71., DOI: 10.1016/j.jbi.2017.05.004

[Report] WissGrid (2010). *WissGrid-Spezifikation: Langzeitarchivierungsdienste,* https://escience.aip.de/wissgrid/publikationen/deliverables/wp3/WissGrid-D3.4.2-lza-dienste-spezifikation.pdf

[Website] ZB MED (2019) *Preservation Policy of ZB MED – Information Centre for Life Sciences*, https://www.zbmed.de/fileadmin/user_upload/ZB_MED_Preservation_Policy_eng.pdf

[Website] ZBW (2018) *Preservation Policy: Guidelines for Digital Preservation at the ZBW,* https://www.zbw.eu/en/about-us/key-activities/digital-preservation/preservation-policy/

[book] American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

[book] Borgman, C. L. (2007). *Scholarship in the digital age: Information, infrastructure, and the Internet.* Cambridge, MA: MIT Press.

[proceedings] Borgman, C. L., Wallis, J. C., & Enyedy, N. (2006). Building digital libraries for scientific data: An exploratory study of data practices in habitat ecology. In J. Gonzalo, C. Thanos, M. F. Verdejo, & R. C. Carrasco (Eds.), *Lecture Notes in Computer Science: Vol. 4172. Research and Advanced Technology for Digital Libraries* (pp. 170–183). doi:10.1007/11863878_15

[report] Consultative Committee for Space Data Systems. (2012). *Reference model for an Open Archival Information System (OAIS)* (Magenta Book CCSDS 650.0-B-1). Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

[report] Rans, J and Whyte, A. (2017). *Using RISE, the Research Infrastructure Self-Evaluation Framework* v.1.1 Edinburgh: Digital Curation Centre. Available online: www.dcc.ac.uk/resources/how-guides

[journal article] Esanu, J., Davidson, J., Ross, S., & Anderson, W. (2004). Selection, appraisal, and retention of digital scientific data: Highlights of an ERPANET/CODATA workshop. *Data Science Journal, 3,* 227–232. Retrieved from http://www.jstage.jst.go.jp/browse/dsj

[online magazine] Rinaldo, C., Warnement, J., Baione, T., Kalfatovic, M. R., & Fraser, S. (2011, July). Retooling special collections digitisation in the age of mass scanning. *Ariadne 67.* Retrieved from http://www.ariadne.ac.uk/issue67/rinaldo-et-al/

[unpublished proceedings] Santini, M. (2004a, January). *A shallow approach to syntactic feature extraction for genre classification.* Paper presented at the Seventh Annual Colloquium for the UK Special Interest Group for Computational Linguistics, Birmingham, UK. Retrieved from ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-02.pdf [report] Santini, M. (2004b). *State-of-the-art on automatic genre identification* (Technical Report ITRI-04-03). Retrieved from Information Technology Research Institute website: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.5.7680

[data set] Waterton, C., Watson, N. & Norton, L. (2013). *Understanding and acting in Loweswater, 2007–2010* [Data set]. Colchester, UK: UK Data Archive. doi:10.5255/UKDA-SN-7359-1