

Towards Narrowing the Curation Gap—Theoretical Considerations and Lessons Learned from Decades of Practice

Journal Article**Author(s):**

Petrus, Ana ; Fischlin, Andreas; Töwe, Matthias 

Publication date:

2016

Permanent link:

<https://doi.org/https://doi.org/10.3929/ethz-a-010667275>

Rights / license:

[Creative Commons Attribution 4.0 International](#)

Originally published in:

ISPRS International Journal of Geo-Information 5(6), <https://doi.org/10.3390/ijgi5060091>

Article

Towards Narrowing the Curation Gap—Theoretical Considerations and Lessons Learned from Decades of Practice

Ana Sesartić ^{1,*}, Andreas Fischlin ² and Matthias Töwe ¹

¹ Digital Curation, ETH-Bibliothek, ETH Zurich, Rämistrasse 101, 8092 Zurich, Switzerland; matthias.toewe@library.ethz.ch

² Terrestrial Systems Ecology, ETH Zurich, Universitätstrasse 16, 8092 Zurich, Switzerland; andreas.fischlin@env.ethz.ch

* Correspondence: ana.sesartic@library.ethz.ch; Tel.: +41-44-632-73-76

Academic Editors: Constanze Curdt, Christian Willmes, Georg Bareth and Wolfgang Kainz

Received: 31 January 2016; Accepted: 3 June 2016; Published: 14 June 2016

Abstract: Research as a digital enterprise has created new, often poorly addressed challenges for the management and curation of research to ensure continuity, transparency, and accountability. There is a common misunderstanding that curation can be considered at a later point in the research cycle or delegated or that it is too burdensome or too expensive due to a lack of efficient tools. This creates a curation gap between research practice and curation needs. We argue that this gap can be narrowed if curators provide attractive support that befits research needs and if researchers consistently manage their work according to generic concepts consistently from the beginning. A rather uniquely long-term case study demonstrates how such concepts have helped to pragmatically implement a research practice intentionally using only minimalist tools for sustained, self-contained archiving since 1989. The paper sketches the concepts underlying three core research activities. (i) handling of research data, (ii) reference management as part of scholarly publishing, and (iii) advancing theories through modelling and simulation. These concepts represent a universally transferable best research practice, while technical details are obviously prone to continuous change. We hope it stimulates researchers to manage research similarly and that curators gain a better understanding of the curation challenges research practice actually faces.

Keywords: curation gap; research data management; digital data curation; data preservation; data lifecycle management; theory lifecycle management; archiving; best practice

1. Introduction

Today's societies expect science to work in a cumulative manner, e.g., [1–4]. While more and more research is performed using digital tools and some of its core activities are already entirely digitized, there is a risk of discrepancy growing between societal expectations and what researchers and the curating institutions actually do.

We understand curation as the management of any kind of data throughout its lifecycle, in order to ensure the preservation of its value over time and its availability for reuse. This data lifecycle point of view dates back to the lifecycle management of data and other information records as used by archives [5,6]. This paper argues that curation of scientific output has traditionally been a cornerstone for progress in science.

For the purpose of this paper, Research Data Management is understood as comprising all activities carried out by an individual or a group of researchers to organize, describe, structure, and store data they produce, gather or use. This includes dealing with data that are in regular use, e.g.,

during the lifetime of a research project and stretches as far as the publishing of results and preserving them for future re-use beyond the scope of the original project. Consequently we consider the management of research literature including all associated meta data as an integral part of Research Data Management, although this information category is often not understood as research data itself. Similarly, in Research Data Management we include theory oriented research that often makes use of complex models and simulation. According to this understanding, Research Data Management comprises core research activities, of which at least some are common to any scientific discipline.

An important trigger for this paper was the imminent retirement of the head of the Terrestrial Systems Ecology group at ETH Zurich. The research of that group has, in part, been on long-term, unique field data resulting from a large research project on the population dynamics of the cyclic larch bud moth in the European Alps, e.g., [7]. Abiotic and biotic field samples of considerable heterogeneous nature have been collected since 1949 using the contemporary state-of-the-art technologies [8]. The technologies have included the use of centralized computing facilities and punch cards for storing and inputting the data for any analysis, which are very different from today's technologies. Experience clearly demonstrates how challenging changes in technology are for long-term research being key for a true understanding of many environmental, notably large-scale phenomena, e.g., [9]. A sophisticated data base [10] including the actual data got lost in its entirety when the computing center was abandoned. Only raw data could be salvaged. As a consequence new, more robust approaches were sought and this experience contributed significantly to the shaping of the archiving technique described in Section 2.

As curatorial institutions, the ETH Library's Digital Curation Team and ETH Zurich University Archives recently discussed the proposed transition of the aforementioned archive to the Public Domain as shown in Figure 1. University Archives are often confronted with data collection requirements when professors retire. The way the Terrestrial Systems Ecology Group organized their archive and cared for it over years was exceptionally well thought out. As described below, it incorporated principles which can also be found in the (younger) OAIS Reference Model, e.g., [11] and thus presented a promising source to transfer data to the ETH Data Archive in the future and publish them.

The main difference of the Terrestrial Systems Ecology Group's approach compared to other groups' approaches is not the use of any sophisticated technical system—there is none—but rather the fact that the group considered what they needed to achieve in their responsibility as researchers, formulated the principles and rules to stick to and enforced these practices. How exactly archiving was done and which decisions were made does not satisfy all theoretical requirements and may not fit other groups easily. However, we chose this as a case study for the following reasons: (i) the archive is remarkably long-lived, it started in 1989; (ii) we wanted to know how the archive survived many technological changes; (iii) we wanted to investigate the value of simply doing what is possible when better solutions are just not available or cannot be afforded; (iv) we were also interested in a minimalist approach, which challenges the merits of large data curation initiatives or the usefulness of comprehensive curatorial solutions or other more sophisticated data archives; (v) finally we wanted to explore the role researchers play and how they can or have to contribute to overcoming the challenges arising from technological changes that put curation at risk in general.

The more scientific activities become digital, the more curation tends to suffer from the fragility of digital data in general and of software erosion in particular, e.g., [12–15]. It seems the discrepancy may become so large, that no curation at all is possible [11,13,16–19], for example because it is prohibitively costly. We call this discrepancy between actual research practice and research content preservation needs the "curation gap", see also [20]. We argue this curation gap is quite critical, since it puts research results at risk of becoming ephemeral, which our case study highlights. Ongoing availability (continuity) of research results is, however, a well known prerequisite for scientific progress, data sharing within and among projects and open science in general, e.g., [15,21].

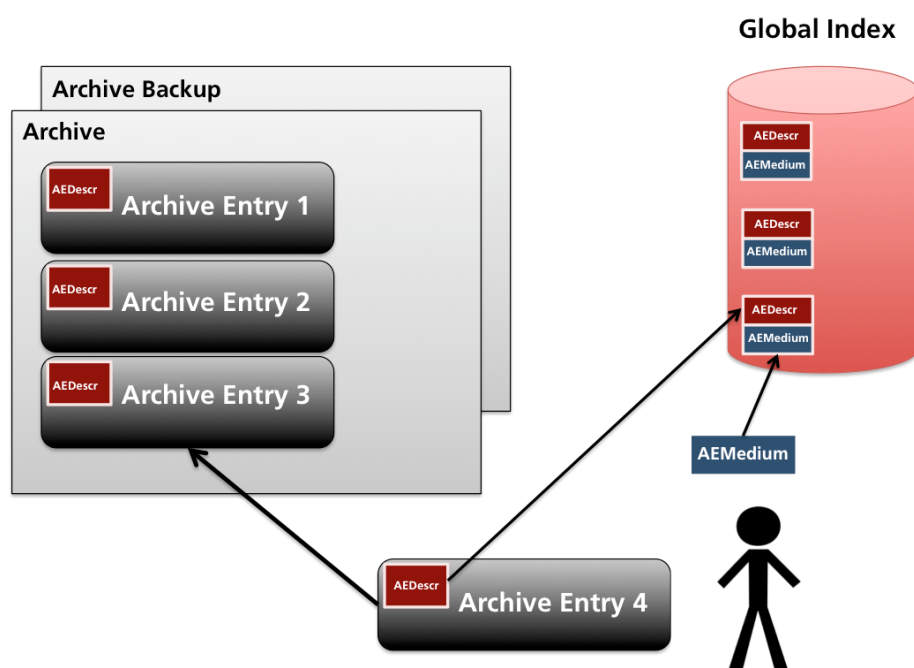


Figure 1. Organization of the Archive of the Terrestrial Systems Ecology Group with two identical repositories (left rectangles) in two different buildings, each containing identical collection of archive entries (grey). Highlighted within each archive entry (AE, grey) are the meta data containing the corresponding archive entry description (AEDescrip, red). For each archive entry the latter (red) is also stored redundantly in a global index residing on a central server (cylinder) for access purposes together with the corresponding media description (AEMedium, blue). All meta data files (red, blue), including the global index, are purposely only plain ASCII text files (examples in Appendix Figures A1 and A2, see text for further explanations).

The curation gap has several perspectives: The first perspective is that of curation itself. The need for digital curation is well acknowledged [22] and many institutions such as universities, funding agencies, and dedicated institutions from the private sector or at the government level are devoting considerable efforts towards it, e.g., [16,23–31]. Initiatives and competence centres such as the Digital Curation Center (www.dcc.ac.uk), DataONE (www.dataone.org), PANGAEA (www.pangaea.de), ANDS (Australian National Data Service, ands.org.au) and a number of others have worked successfully to establish services and best practices. However, in many regions and subject areas these efforts often only reach a minority of researchers. Moreover, curation itself currently faces many uncertainties, not the least due to unknown costs [16,17,23,32–34], plus several barriers and other difficulties [6,18,20,22,35,36]. There is also a particularly problematic interaction with Open Data. In disciplines with little tradition of data exchange, there are hardly any incentives to manage one's own data in a way that facilitates sharing, e.g., [21,34,37], let alone immediate sharing. Once badly managed, it can take major additional effort to share data or may even become impossible.

This leads to a second perspective, the actual research practice. How are data and workflows managed, in particular with respect to digitization including any workflows involving digital data?

Some researchers and other stakeholders have argued that the lack of available tools and infrastructures obstructs digital data curation [26]. Many argue the situation is best improved, by more and better education of curating professionals and researchers, facilitating mutual understanding [38–41] and a few also call for targeted training of individual researchers with respect to digital data curation, e.g., [37,42]. Some researchers ask for better support, including that from their home institutions as well as from national or international infrastructures [19]. Others have made efforts to provide such a support and currently an active community of librarians, archivists

and similar information experts are busy providing the necessary means, e.g., [16,23–31]. But will this suffice to narrow the seemingly widening curation gap? With a growing number of initiatives and institutions engaging in research data management, one might expect the gap to be closing. Our experience, including from our case study, calls this expectation into questions. The pace of development in research itself is currently so fast that this appears to be far from assured.

Researchers themselves face challenges, have particular incentives, and work in a particular environment. All of these factors need to be well understood in order to address the issues of digital data curation appropriately, e.g., [21,43]. Additionally, it is important to note that challenges considerably differ among the types of research output, for instance output may come in the form of raw or processed data, or models versus well defined and widely institutionally supported outputs such as scientific publications. Scientific journal articles are created in a standardized way with the purpose of becoming part of the formally published record of science. Published works are expected to remain available in the long term and this is also true for electronic publications. Yet, electronic publications face particular additional challenges in terms of acquisition, usage, e.g., [35,44,45], and preservation, e.g., [46]. Nevertheless, we believe that memory institutions such as libraries are in a reasonable position to warrant the curation of well defined content such as scientific publications regardless of their form, even if they only obtain it after its production is finished.

However, this is not the case for research data and research know-how, including large and complex models, that are typically not formally published. Today they are at best made available in the form of supplementary material, but even then typically little is cared for in terms of their curation. Many publishers disclose any responsibility for the readability and usability of the supplementary material while authors believe to have it secured, only because it is stored at the publisher's web site.

For the less formalized digital objects such, which make up the majority of today's research data, we argue that there is a critically relevant discrepancy between the offers provided by curation experts and what most researchers are currently practicing, *i.e.*, there is a critical curation gap, which may even be widening. The following reasons are likely to continue contributing to this regrettable trend:

- The issue is not, that available training and tools do not support Research Data Management, e.g., [47], but rather that by the time they find their way to researchers they might already have moved on to more sophisticated methods requiring other tools. It might therefore be necessary to focus on very basic concepts independent of specific tools instead of trying to keep up with the pace of technological change.
- Curators often lack precise understanding of the actual practice of research and consequently risk being ignored by the research community in their daily practice and vice versa A good mutual understanding would be needed to prevent obstacles that can no longer be overcome once research projects have ended, e.g., [21,42,43].
- Research practices strongly depend on a specific research community's methods, traditions, and standards, limiting the influence of "outsiders" such as curators.
- Researchers face a competitive environment and often have little if any incentives for managing their research data and other research activities, including cooperation among researchers, in such a way that would support curation.
- Total costs (time, resources) of curation may impede digital data curation and even become exorbitantly high.

Likely institutional responses to these trends include the launching of efforts for minimizing the total cost of digital data curation in terms of time and resources as expended by all involved and reforming research environments towards an enhanced quality of research data and results in terms of their "curatability". However, such reforms tend to progress slowly, in particular vis-a-vis current rates by which digitization is advancing.

We argue that a better understanding of the issues by the involved communities can help immediately. In this paper we present ready pragmatic solutions. However, they require that

curation needs are well addressed instead of ignored. Here we draw from our case study an experience spanning over almost three decades of Research Data Management efforts, including a simple archiving concept, as developed and followed through by the Terrestrial Systems Ecology research group at ETH Zurich.

In this paper, we first describe the archive and then the basic concepts as developed and applied by that research unit by focusing on those parts of research that are common in any type of research, ranging from natural to social science and the humanities. We discuss the following research activities, each treated in its own section: (i) handling of research data, (ii) publishing, including reference management, (iii) management of theories, notably as encapsulated in models. Following a narrow definition of the research data, it might be surprising to discuss literature and reference management at some length at the same time. However, literature and reference management should be regarded as an integral part of the research process, which improves the quality of research. The challenge here is less long term, but rather the efficient management of members of a group and their parallel or successive projects. The concepts are presented in an illustrative manner, yet with sufficient detail, e.g., by explaining how commonly available tools were used to make the concepts—partly quite sophisticated—easier to understand. Some technical details that may be of particular interest for some readers are contained in appendices (one for each discussed research activity). We then discuss some lessons learned from the experiences described and what it means in the current situation of a widening curation gap. We hope that this will enable progress in all involved communities, including researchers, curators and tool makers.

2. An Intentionally Minimalist Archiving Concept

For several years, ETH Zurich regulations have stipulated that all data and models created during employment at ETH Zurich belong to the institution and not to the individual authors, with the exception of their non-negotiable right as creators. ETH Zurich research policies also state that all research needs to be traceable to ensure full transparency and proper conduct. Archiving has thus been a duty for each employed researcher at ETH Zurich for some time. However, the policies leave it to the principal investigators to define the requirements according to common practice in their domain and to address them by the means they consider appropriate. This leads to a wide variety of approaches being employed, usually at the level of research groups. Although no reliable or up-to-date overview of those solutions exists, experience from the authors, who work at the ETH Library, indicates that mere storage of uncommented folder structures or disk images is far from uncommon, in particular, when data is not considered as part of any long-term research. More sophisticated solutions also exist, but they are not the rule and their management over generations of doctoral students can be a challenge. Although already formed in 1988 and therefore probably before most of the above mentioned requirements came into force, the Terrestrial Systems Ecology Group faced particular challenges in managing its research (see also Section 1).

Its interdisciplinary research depended heavily on complex simulation models and diverse data sources. The leader's previous experiences of pioneering the field in the 1970s using centralized computing facilities demonstrated the need for robust Research Data Management that could smoothly survive the rapid transitions information technologies experienced during the 1980s, not the least due to the onset of personal computing, e.g., [48].

The most urgent need for data management came from the fact that the group pursued field research, *i.e.*, continued the larch bud moth project, which had begun with extensive field measurements along the entire European Alps in 1949 [7,8]. Since its very beginning that project continuously sought to apply the then most modern technologies, including computing. Over the course of the research collected data were stored on many media including punch cards, paper tape, magnetic reel tapes, *etc.* Computing and storage facilities were still extremely limited in the 1970s and early 1980s. Nevertheless, one of the authors, together with co-workers and data base specialists, developed a state-of-the-art data base late in the 1970s to hold all types of larch bud moth

research data from the field and laboratory (LAWIDAT as a DDLML-INFOSYS data base, [8,10]). The purpose was to ensure long term storage of all data and continuous data entry as soon as samples were collected. It also aimed to improve on data quality in terms of consistency, immediate error checking, and the entering of meta data, in order to provide easy access to data for analysis and modelling to all involved researchers, and to enable more comprehensive data analysis.

However, this data base solution was far from sustainable in terms of its manpower requirement. By the time ETH Zurich's computing centre was abandoned in the late 1980s, it had become prohibitively costly to transfer that data base system to a modern host. Software erosion had also made it impossible to retrieve all data in an orderly fashion and the precious data could mostly only be salvaged in their raw form.

This experience influenced the decision to look for a new system that was intentionally minimalist, while ensuring economy and persistence. The approach was to minimize dependencies from sophisticated technologies requiring costly maintenance, yet to manage the data according to some fundamental principles that form a system comparable to the functions of a data base system.

The resulting Systems Ecology Archive was made in 1990 according to the following concept.

Self-containment The main purpose was to ensure the long-term use of the data for research occurring over decades. The archive also needed to be self-contained, *i.e.*, not only observational data and measurements (research data in the narrow sense), but also all involved software such as models, applications, and operating systems were archived. Even hardware was preserved when necessary. Copyright issues in conflict with self-containment were avoided by limiting access to the archive to the research group.

Archive entries At the end of each project or a well-defined project phase, an archive entry was to be made (Figure 1, AEDescr). A research project was only considered really ended upon the completion of its last archive entry. Each entry comes with a description and is identified by a globally unique title and forms part of the archive. Thus, all of the archive's meta data could anytime be reconstructed from the archive, also contributing to the self-containment of the archive.

Meta data The meta data were split into the describing part and the media part (Figure 1, AEDescr *vs.* AEMedium; illustrative examples in Appendix A.1, Figure A1 *vs.* Figure A2), where the latter were kept outside the archive to be updated during regular maintenance, *e.g.*, as storage media would age and require copying. Archive entries were only added and were not allowed to change in any way, regardless whether the storage media would in principle allow for overwriting or deletion. Updating of an archive entry would have to be accomplished by rearchiving an updated version. Finally data base software was not used on purpose. All these design choices served to minimize archive maintenance.

Formats All meta data were only stored in simple ASCII encoded text files, while the actual data were typically archived in original formats. However, most critical parts, *e.g.*, a dissertation text, were also redundantly archived as rtf, plus text files, or spreadsheets containing precious data, again redundantly, such as SYLK, plus text files (for details see Appendix A.1).

Storage media Access to the individual archive entries depends on the file system in use, which may follow standards or not, depending on the storage media. *E.g.*, magneto-optical disks, favoured due to their expected long lifetime of 50 years, were pragmatically formatted with the then used computer platform, possibly requiring archive maintenance at some point in time by moving affected entries to new media. CDs or DVDs were burnt according to ISO standards.

Access and use To facilitate the retrieval, a Global Index (Figure 1, red cylinder), *i.e.*, a global collection of archive entry descriptions, was stored redundantly outside the archive on a central file server that every researcher could access. In this ASCII text file all archive entry descriptions, extended by their media descriptions, were accumulated in chronological order. The search for a particular archive entry depends on the searching capabilities of a text editor or a command line tool such as *grep*. Since the primary role of the Systems Ecology Archive was to enable the internal use of data for research purposes, access was restricted to research team members only.

Further details on the specific characteristics of this archive, including examples of meta data, workflows for preparing archive entries plus actually used rules for access and maintenance, are contained in Appendix A.1.

3. Handling of Research Data

All handling of research data by researchers themselves during a research project can be understood as phases of one continuum with respect to preservation and curation as discussed by Treloar and Harboe-Ree [49] and shown in Figure 2.

How well data can be preserved over time and if it can be re-used depends to a large extent on the measures taken at the time of data production. This dependency is even more pronounced in those cases where curation is delegated to specialized institutions, which often only come into play at the “end of the pipe-line” of data production. Not surprisingly, the influence of such institutions is rather limited. At best some less direct influence can be gained by curators via policies, guidelines, and training. Issues regularly encountered concern missing meta data and context documentation, missing information on file formats (properties, dependencies, related software), lost or corrupted data files, and missing or unclear information on legal aspects (rights to access and use of data, as well as third party rights), e.g., [13,26].

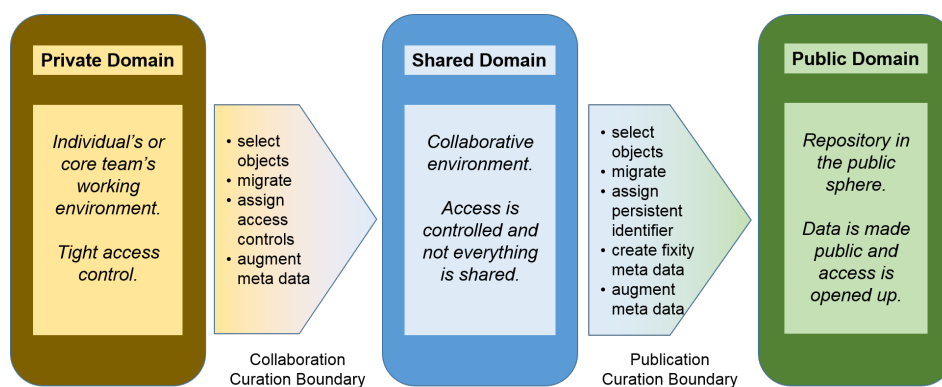


Figure 2. Transitions between curation domains along the research data lifecycle, based on the Data Curation Continuum figure by Andrew Treloar [50].

As there is often a time lag of several years (or even decades) between the onset of data production or collection and the decision to publish and preserve data, there is a risk that these issues do not receive the necessary attention at the time the information gaps could still be closed easily.

However, information needed to identify, retrieve, interpret and use data, is of course crucial for any meaningful curation. Information that should be collected while identification and retrieval is still possible includes information on technical dependencies and characteristics, descriptions of procedures and tools, legal information in cases where use is restricted, and a full documentation of the scientific context, in particular underlying theories, models, algorithms, and publications. As trivial as this may be, our experience shows that the following is often forgotten: most of this information can only be provided by researchers themselves, in particular any information which is needed to facilitate a scientifically sound re-use of data in the future. Relationships may be quite asymmetric, e.g., Rule 5 of [42], for example, while a researcher may consider the relationship between a scientific publication and the underlying data to be obvious and safe, since published, our experience shows that during a late curation it is often impossible to connect those data to the publications it was used for unless the researcher has added explicit reference to the data and all related publications (this n:n relationship is very rarely properly handled unless curation aspects are considered from the very beginning and policed).

A few principles for the management of data collections need to be observed, not only by curators but also by the researchers.

Firstly, one needs to distinguish data by their quality, *i.e.*, raw *vs.* processed. It could be tempting to consider raw data as an easy candidate for curatorial tasks because it has not undergone further tool-dependent transformations which would need to be documented. Unfortunately, this is rarely meaningful, since the quality of raw data usually prevents their as-is use, since data need to be checked and possibly transformed first, which is best done by the producer.

Secondly, processed data represent the product of the application of procedures or algorithms to raw or less processed data. While this may seem trivial, massive, yet often overlooked, consequences arise with respect to documentation of such processing. Unless well documented (best handled by archiving detailed descriptions of the used procedures and even better by also including the used programs themselves) data exchange or re-use, *e.g.*, while appending some new raw data, become either impossible, meaningless, or otherwise scientifically highly questionable.

Thirdly the kind of information required for such purposes has been comprehensively described in the Reference Model for an Open Archival Information System (OAIS) [11] (full text mostly identical with ISO Standard 14721:2012 [51]). In that model any data collection can first be seen as a Data Object. This model emphasizes that a mere Data Object can only be rendered or meaningfully used if it is accompanied by appropriate Representation Information (pp. 2–4, [11]). If only the data object is available, rendering does not work and the wanted Information Object remains inaccessible (Figure 3, top) and archiving is not worth the effort. Data Object and Representation Information form the Content Object together, which can then be processed to yield the wanted Information Object (Figure 3, bottom).

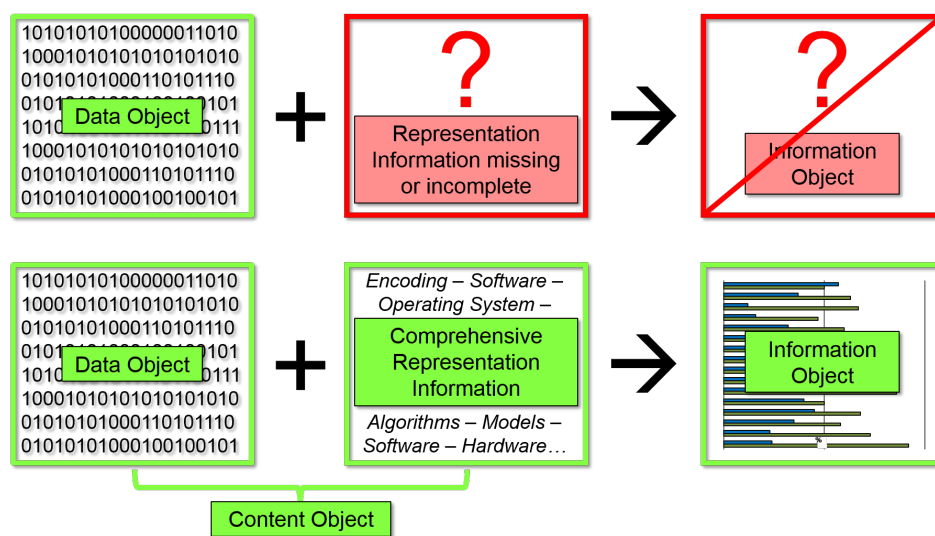


Figure 3. The information contained in a Content Object can only be accessed if both the Data Object itself and any Representation Information it relies on are available.

Representation Information must cover the whole chain of dependencies a Data Object relies on, in particular specific software needed for rendering and using it, the operating system such software relies on, and even the hardware needed for running the operating system. In principle, Representation Information would have to package software and operating systems in the exact version available at the time for archiving as part of the Content Object.

As this would be a highly redundant approach, it would be more useful and efficient to reliably collect these tools in a common repository and thus only have to document and persistently reference them in the Representation Information. While this approach of shared resources can already

be implemented at the level of research groups, it would be preferable for efficiency reasons to organize this at the level of entire institutions or even international communities with the support of institutional custodians. However, current institutional support for such approaches is minimal.

Fortunately, the better defined—and therefore usually more uniform in its specialization—any given scientific community is, the further the common knowledge, including used tools, can be extended. Significant savings by widely omitting exhaustive Representation Information from the archiving would then become possible. Knowing that all components needed for rendering are reliably curated by the community elsewhere, a mere referencing of those standards would suffice.

Another dimension worth-considering in this context is the period of time for which a Content Object must remain fully usable. The shorter this period, the more assumptions can be made about the general availability of the software outside the archive itself.

No matter how this is eventually approached, it is obvious that it will not be possible to always document the complete chain of dependencies in a universally unambiguous sense, given the fact that each element of the chain would have to be considered a Data Object in its own right. Technical and resource restrictions render this infeasible.

A pragmatic approach towards addressing these issues is to take at least minimal organizational measures at the level of a research group. At that level critical dependencies of data and their processing, e.g., specific tool requirements, are known the best. The simple archiving concept from the Systems Ecology Archive (Section 2) follows such an approach by supporting mutual cross-referencing and dependency listing of archive entries (Figure A1). Some often shared components, such as operating systems, application programs, and discipline specific tools, were archived separately with different entries, not only reducing redundancy and minimizing archiving efforts, but also significantly enhancing the self-containment of the archive in terms of the reusability of the archived data. Our case study demonstrates that today's hardware emulators combined with legacy tools retrieved from the archive supported a surprisingly long-term data, tools, and model use, thereby significantly reducing the damaging effects of software erosion.

Another important element of the approach taken in the Terrestrial Systems Ecology Group was to process raw data quickly and bring them into a form that contains a minimum of meta data facilitating reuse. This made it possible to archive such data sets any time, since in processed form the repository may hold at least references to entities, e.g., tools, reports, or publications, contained in other archive entries, that contain the needed representation information (see also Section 2, Figure 1).

To this end so-called Data Frames were used, a data storage format, which offers, among many other advantages, the ability to store processed data together with meta data (Figure 4, details in Appendix A.1; Data Frames were primarily developed for modelling and simulation purposes as discussed in Section 5). Other formats can serve similar purposes by reserving separate sections, e.g., at the beginning or end of a large data file, for holding meta data. However, other formats may have considerable restrictions compared to the Data Frames written in a formally defined LL(1) language, e.g., [52] supporting the functionality of an entire data base system.

Storing meta data together in the same file as the data themselves facilitates curation. The rendering information may then be fully contained in the meta data (Figure 4, bottom) or at least referenced (Figure 4, top). In the latter case the Information Object can only be rendered if both the referencing as well as the referenced archive entries are available (*cf.* Section 2 and Appendix A.1).

The origins of the case study discussed in this paper considerably pre-date the creation of the OAIS-model. While the OAIS-model and its associated complex tools were lacking, clear concepts compensated for this lack and empowered an archiving approach of a comparable power to the OAIS-model. We argue this experience should still be understood today as an encouragement to do what can be done right now, instead of deferring a data management facilitating curation to an uncertain future while waiting for a comprehensive technical solution or otherwise ideal service. Services that would be optimally convenient for the user might never evolve or only very late and when the current research needs are finally matched at some point in the future, they may already

have grown or changed. This argument also becomes stronger the more economical constraints restrict the involved research or the subsequent maintenance of the once chosen curation approach.

```

...
DATAFRAME SiteEdaphics; MODEL = ForClim; DATA:
(* see Fischlin et al., 1995. doi:10.1016/0269-7491(94)P4158-K Fi04dg *)
(*-----*)
Site      Bucketsize  SiteID  ;
(*-----*)
"Bern"    30.0           333333  ;
"Bever S" 20.0           222222  ;
"Davos"   25.0           666666  ;
...
END SiteEdaphics;

(*
Tree species for biome region 'Central Europe'

Access in models species names by value definitions similar to
"ForClim.Fsil.Scientific_name" or "ForClim.Psyl.Common_name".

References: Heitz et al., 1990. Schul- und Exkursionsflora fuer die
Schweiz mit Beruecksichtigung der Grenzgebiete. Schwabe & Co., Basel,
ed. 19, 659. He196
*)

DATAFRAME TreeSpecies; MODEL = ForClim; KEYCOLUMN = SpecIdent; DATA:
(*-----*)
SpecIdent  Scientific_name  Common_name  ;
(*-----*)
Fsil       "Fagus silvatica L."  "European beech"  ;
Aalb       "Abies alba Mill."   "European silver fir" ;
Psyl       "Pinus ..."
...
END TreeSpecies;
...

```

Figure 4. Example of data stored as so-called Data Frames following the DTF syntax (cf. Figure A7 in Appendix A.3). The DTF syntax allows meta data to be entered, intervening anywhere in the file as comments. These can be used for adding minimal meta data to the actual data (**top**), i.e., merely referencing other critical components worth-archiving, such as a separately to archive publication (cf. Section 4.1), or rather lengthy self-contained documentations (**bottom**).

4. Publishing and Literature Management

4.1. Scholarly Publishing

The publication process generates a defined product in the form of a formal publication, but in both a scientific and curatorial respect, this forms only the tip of the iceberg. The formal publication presents the results of much underlying research in an intentionally condensed form while most of the data and materials actually used remain unpublished.

At best, the data and collected materials are made partly available in the article's supplementary material. However, the quality of this material is often quite questionable, for example, many publishers offer supplementary material only on an as-is basis delegating any responsibility for content and readability to the authors, while enforcing no standards for such material. Moreover, we observe that authors tend to underestimate software erosion by offering supplementary material that is accessible only with proprietary software, so that the supplementary material can become unreadable at any time while the main article is still readable.

Fortunately, more and more funders and journal editors are pushing for the open publication of publicly funded research. However, even if this would be widely done, there would still be many unpublished data that need to be preserved according to what we call good scientific practice.

For both purposes—making relevant data available to others and keeping it for reasons of accountability—the experience from our case study indicates that it may be helpful to treat each

publication as a project on its own. According to that concept, at the end of each project comes the archiving (*cf.* Section 2 and Appendix A.1). The archiving of a publication is based on the principle of packaging all data, models, and tools that were used to create the publication, including, e.g., scripts that were needed to process the data and/or create graphics. The guiding principles are thus the same as for the handling of research data as presented in previous section (*cf.* Section 3).

4.2. Literature Management

The following section deals solely with literature data (bibliographic meta data and full texts) and intentionally excludes research data *sensu stricto* treated elsewhere (Section 3). Notably, however, data papers are becoming more and more common and hereby blur our distinction.

Literature management understood in this way encompasses the entire research process ranging from reading background information up to the automatic generation of the reference list while authoring publications, while always containing a significant part personal to the individual researcher, e.g., [53]. It should therefore be considered as an omnipresent, integral part of Research Data Management as we define it for the purpose of this paper. Since researchers typically accumulate several thousands of references and papers during their careers, an efficient management is only profitable if it handles this information economically across research projects, publications, the researchers' CVs and web site plus its sharing with collaborators, peers, and co-authors within and beyond ongoing research projects.

Good literature management is still a challenge, as, contrary to what many may assume, current trends are contributing little towards narrowing the curation gap. This topic is also a case in point showing that the curation gap is not only the responsibility of the research side, but also requires improvements from the side of Library and Information Science and all its related services and tools. The entire life cycle of literature data creates considerable challenges for scientists in their daily research, and needs to be well understood to truly progress in narrowing the curation gap. While we have the least concerns for the curation of scientific publications themselves—as today they are being well handled by publishers and libraries across the globe—this is not the case for the publications' meta data and the annotations typically done by researchers on personal copies.

There are numerous reasons why we also have a curation gap in this area. Here we mention a few: (i) Researchers typically use several different tools to write articles (e.g., Microsoft Word, L^AT_EX) making it necessary to use several reference manager tools (EndNote, Mendeley, BibT_EX, *etc.* (The listed reference manager tools are listed and discussed here rather arbitrarily and no favouring nor criticism of any of these products is intended. We shortlisted here merely a few of the more popular ones as representative for many others.)). (ii) The use of such tools may be prescribed by publishers or is decided by an author team and is thus beyond the full control of the individual researcher. (iii) Many of today's most popular reference manager tools are not based on a solid data base foundation by violating basics of data base theory (lack of a primary data base key, e.g., EndNote, Mendeley) or considering basically the researcher to be the only author of an article (e.g., EndNote) or by having insufficient provisions for reliable import, export, and updating of records (e.g., Mendeley), let alone the curation of the involved data. (iv) More or less competing meta data repositories for scientific publications are offered by many, mostly internet based services. They are typically ahead of the individual researcher and economic reasons make it necessary to make good use of such services. However, these services often employ proprietary techniques preventing the flexible use and reuse of the involved meta data within a situation where several writing techniques are forced onto the researcher. In summary, there is no one size fits all solution available and it is unlikely to come in any foreseeable future. As a consequence, researchers have to continuously adapt their writing techniques and require a flexibility that is generally far beyond what is given by any of today's services, let alone the relevant software tools.

Here we describe a system that has been successfully implemented and used at the Terrestrial Systems Ecology Group at ETH Zurich in an attempt to address such issues. The system has evolved

and survived multiple technological changes and has been in daily use for nearly three decades. It can serve as a successful model of how to implement robust literature management within a research group and how to pragmatically overcome some of the aforementioned challenges.

The goals of the system proposed here were built around following design principles. While attempting to share the literature collection among research team members, each individual team member should be given maximum freedom to maintain a very personal collection. While some see these goals as extending conflicting objectives, the following design rules and explanations should make the choices taken understandable.

The proposed literature system for personal collections of publications is illustrated in Figure 5. This system supports the sharing of bibliographic meta data and full texts of publications, possibly annotated, within a team of researchers, while maintaining the basic personal ownership of the collection by each participating individual researcher x .

The system works whether a researcher x is connected to the internet or not and maintains the integrity of the data, despite the complexity of its topology supporting the distribution of the bibliographic meta data and full texts of publications over multiple users and multiple devices (see also Appendix A.2). This gives the individual researcher x the feel of working with a personal literature collection at all times, while also supporting collaboration, including by the sharing of at least meta data on the literature in use within the team.

Moreover and quite importantly, this system supports not only the management of meta data, but also the maintenance of a personal publication collection, e.g., in the form of a collection of PDFs, the reading of those publications, plus annotating those publications (see Appendix A.2 for technical details on the actual techniques and tools used). This concept attempts to make the system as attractive as possible from the perspective of the individual researcher, thus significantly increasing the probability of its use, as nearly three decades of experience indicatively confirms.

Meta data records containing bibliographic references of any scientific work are exchanged among users of the research team only via the central data base *LitCentral* (star-like topology). This happens by synchronizing records from the owner via *LitCentral* to another user's personal data base, e.g., from *LitMY_a* as used by researcher a to *LitMY_b* as used by researcher b . This is assumed to happen asynchronously, *i.e.*, whenever any of the involved researchers x connects to the central data base *LitCentral*.

Consistency among researchers is warranted by associating particular permissions to the records. The individual researcher, e.g., a , has write permission (green) on all her records, but only as long as those records reside on a device owned by that user. Note, domains A, B, \dots are user specific (1:1 relationship), yet are not device bound so that an individual researcher can use as many devices as she likes (Figure 5 illustrates this for domain A with a tablet computer and a smart phone, for domain B with a laptop and a smart phone). Every record is owned at all times by only one researcher (1:1 relationship) and must have a globally unique key referencing a particular scientific work (1:1 relationship) to warrant consistency throughout the system at all times. Unfortunately, this property is where most of today's, e.g., cloud based solutions (e.g., Mendeley), fundamentally fail.

For illustrative purposes Figure 5 uses keys that consist of a number and the letter denoting the record owner, *i.e.*, researcher x . Any duplicates of meta data records (1:n relationship) may be distributed throughout the research team, however, only with read-only permissions (in Figure 5 sections are coloured red if the record copy resides in a domain other than that of the record owner).

If the key also contains the owner of the record, any key assigned within any given domain X that is unique within that domain X , is then automatically also globally unique. This greatly facilitates an independent assignment of keys, but comes with the disadvantage of having redundancy within the entire system by allowing several researchers to assign different keys to the very same scientific work.

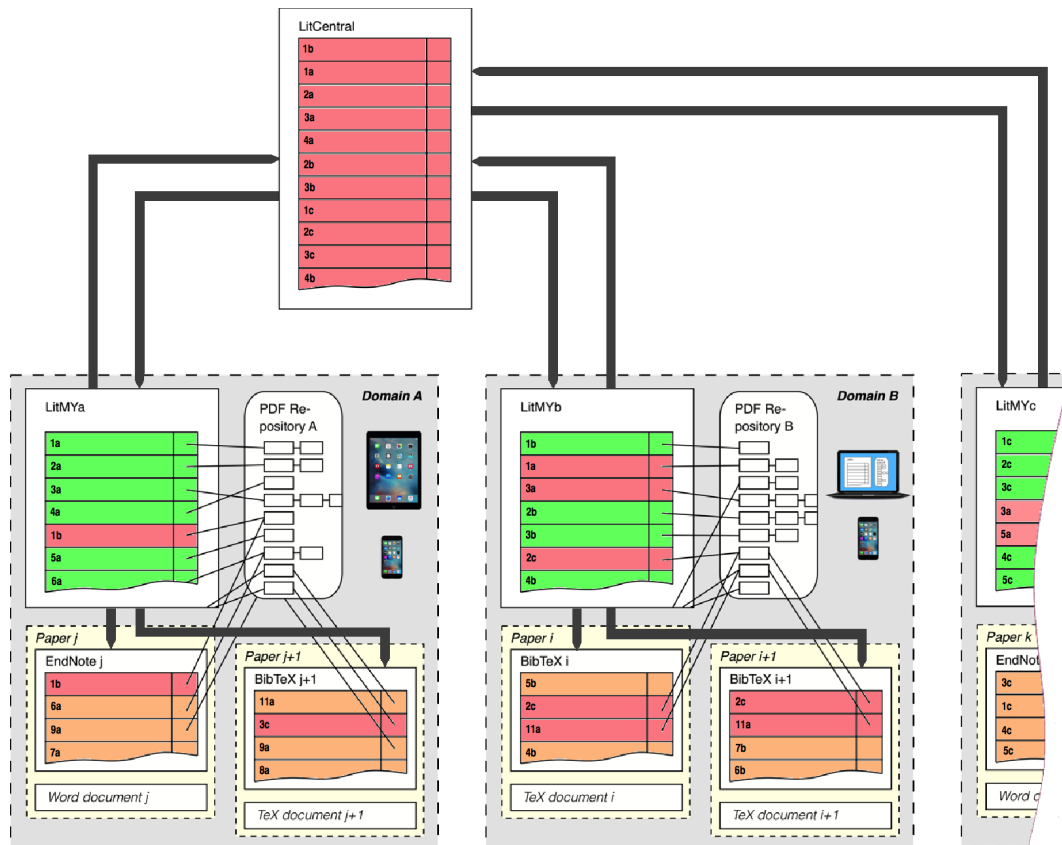


Figure 5. Sharing via a central data base *LitCentral* meta data on personal publication collections (*LitMY_a*, *LitMY_b* ...) among researchers *a*, *b*, ..., belonging to the same research team. A record referencing a scientific work, e.g., a journal article, can exist in multiple copies distributed throughout the system. However, only records in green are master records and have write permission, whereas any redundant copy used outside of the personal data base system, say *LitMY_a* as used by researcher *a*, is a read-only copy (colored red if in domain of another researcher, orange if used within the researcher's own domain as an excerpt for a specific project such as paper project). For convenient reading every researcher typically links records at all times to a personal document repository, typically a PDF repository, while possibly using multiple devices within the personal domain, say domain *A* (here we assume devices can be well synchronized within the respective domain) (See text for further explanations).

Yet, as our experience shows, there are techniques to minimize the risk of such redundancy, e.g., by having convenient tools to check for the presence of a record already referencing a particular scientific work. With the spread of digital object identifiers this disadvantage could be further remedied. Figure 6 illustrates the corresponding rules to observe when entering new records into the system. However, note, the design principle proposed here makes it attractive for every participant to use such facilities, since it reduces users' workload, enabling them to merely share a record instead of going through the trouble of re-entering it.

Any possible updates of the core data of every record owned by user *x* will overwrite the corresponding copy in *LitCentral* as soon as researcher *x* synchronizes her *LitMY_x* with the central data base while merging any non-core data such as personal annotations from all record users.

Any other user, say user *y*, possibly using some of those records as well, will then overwrite his copies within domain *Y* as soon as he synchronizes his *LitMY_y* with the central data base *LitCentral*. This means asynchronous synchronization of the data among the various copies is always unidirectional from the master record within the owner's system and has to propagate throughout the system to other devices within those domains only as users access the central data base *LitCentral*.

Again, this gives full control to all participants when the sharing of bibliographic data and full texts takes place, while keeping the necessary techniques simple and ensuring data consistency at all times.

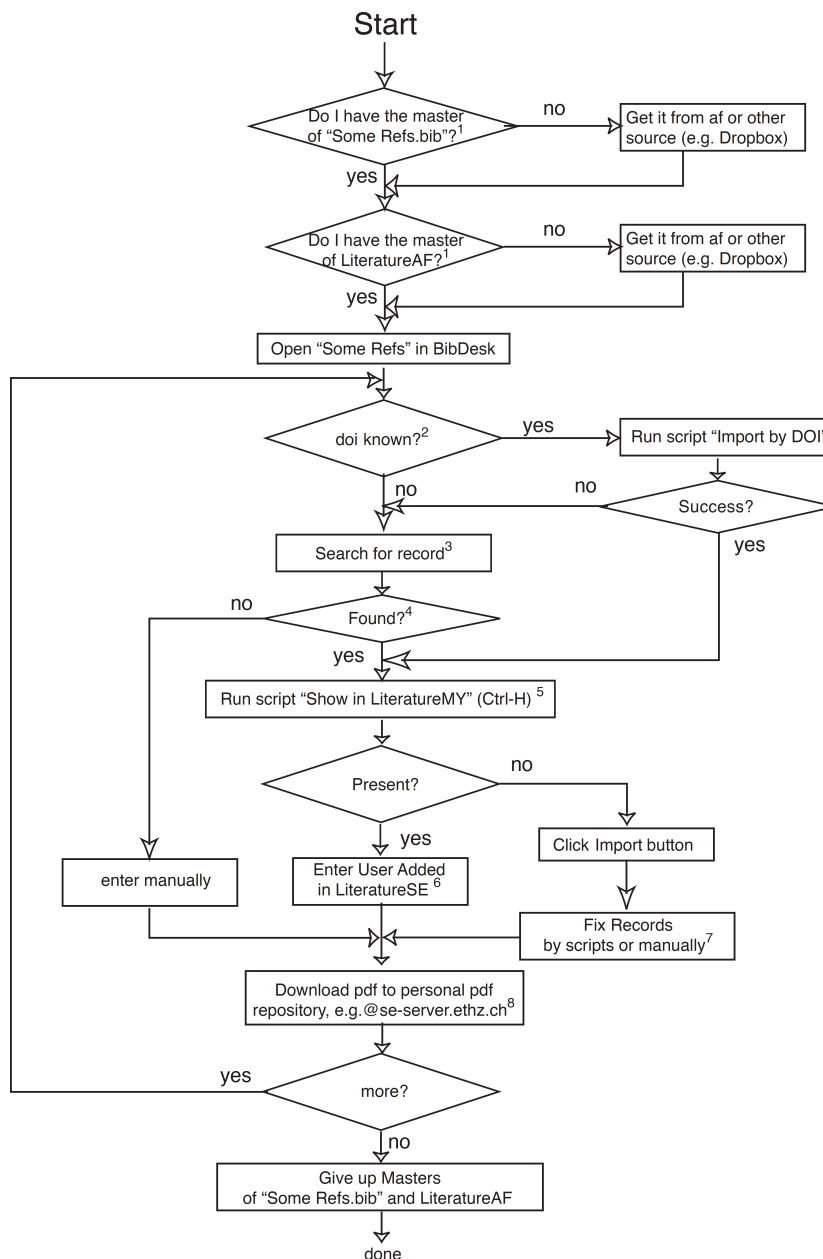


Figure 6. Flowchart illustrating simple rules researcher x should observe when adding new scientific works to the system proposed here (cf. Figure 5). Here it is assumed that this work is done within domain A of researcher a (abbreviated by “AF”) using the open-source application *BibDesk* and for *LitAF* a *FileMaker* data base (See text for further explanations).

Experience of almost three decades, having implemented this concept by using pragmatically various components of widely used tools such as *EndNote* and *BibTeX* among others while authoring scientific works using *Microsoft Office Word* or \LaTeX , has shown that such a system works very well, remains efficient and supports research—in particular collaborative research—in many ways. Notably, this time-tested system is expected to support the use of existing tools in a pragmatic and flexible manner. This enables the individual researcher to switch tools on the fly, and as needed, when collaboration changes or merely when another scientific journal imposes the use of other specific tools

on the researcher (e.g., *Word vs. L^AT_EX*). If tools change more permanently or other tools need to be incorporated, the system can co-evolve and even partly compensate for the many shortcomings and deficiencies that today's more popular applications show when it comes to maintaining consistency.

However, this system is not without disadvantages. E.g. every researcher x only needs to have, in principle, at all times a single master copy corresponding to a particular publication within her domain X . Exceptions to the aforementioned requirement ought only to be tolerated temporarily, *i.e.*, while entering a new publication and in a transitory state of the system (e.g., Figure 6). Yet, this requirement can only be "enforced" by the discipline of researcher x , which creates a real risk of inconsistency, despite remaining entirely within the single domain X controlled only by single researcher x . This limitation is largely due to the pragmatic nature of our concept, *i.e.*, allowing the incorporation of popular tools, despite their flaws, *i.e.*, not being designed for cooperation with tools from the competition.

Being forced by changing collaborative research to use, e.g., multiple "Reference Managers" also adds complexity, another hurdle that may be perceived by researchers as a considerable deterrent for sticking to our concept. Yet, the aforementioned requirement is critical, must not be relaxed, and requires strict discipline from researcher x to avoid any shortcuts by only updating green records, and never orange ones (*cf.* Figure 5), regardless of where a deficiency in some meta data, e.g., a typo, was detected. Updated green records need to be propagated within domain X by transferring the updated green records from central *LitMYx* to any other data base storing a project specific excerpt of the orange records. The risk of inconsistency can again be minimized by offering convenient transfer tools that can efficiently accomplish such propagations conveniently (Appendix A.2, Figures A4 and A5).

Despite certain disadvantages that come with our approach and numerous deficiencies of today's "Reference Managers" (*cf.* Appendix A.2 for details), which are not helpful, we have been able to develop pragmatic, yet very successful solutions. We needed only to follow clear concepts that provided the critical foundation to warrant consistency within the system and consequently reliability (Figures 5 and 6), one of the most relevant elements to user-friendliness.

Moreover, our concept seemed also to be largely in accordance with individual researcher's motivation and incentives—notably by supporting personal use—while nevertheless supporting collaboration and sharing of literature data. Our conceptual design (Figures 5 and 6) appeared essential for avoiding issues and minimizes risks of inconsistency. It can also be generalized independently of particular tools as long as the following requirements are satisfied:

- each participant can be globally uniquely denoted (e.g., by using the ORCID unique researcher identification).
- each publication is also globally uniquely denoted by a main key (e.g., DOI) within the distributed data base system that also identifies the owner
- all files, e.g., PDFs, associated with a given publication are named so that the file name contains the main key or allows to derive the main key to be derived

A central registration server would satisfy the first requirement, even for very large numbers of researchers. For requirement two a good candidate for a main key would be to construct main keys from the widely used DOI (Digital Object Identifier) of the referenced scientific work concatenated with the user ID as resulting from requirement one (assumed all involved works have a DOI). Requirement three is in its simplest form using the main key for the file name, resulting from satisfying requirement two.

However, the current generation of DOIs is uses characters that are technically in considerable conflict with today's file systems and "Reference Managers" coding schemes. Together with the fact that far from all publications having assigned a globally unique DOI, this all speaks against the here discussed straightforward general solution. Considerable revision of widely used schemes and software tools would be required to satisfy all three requirements well, to then offer, e.g., a cloud based system with even more convenience than our system currently can accomplish.

Finally, while every professional data base system is fit to accomplish the needed tasks and could in principle support our system, the citing of references while authoring publications is not at all supported by them. What would be needed is also the latter functionality, which belongs to the typical strengths of the “Reference Managers”. However, the latter have considerable weaknesses in terms of data base functionality, another argument clearly speaking for a solution similar to the one we presented.

It appears that most of today’s “Reference Managers” fall short in accomplishing the fundamental data base functionality as actually needed and/or contribute too little to improving the situation by serving the actual research needs as described initially. Our approach, however, addressed all relevant issues. It offers the functionality of a fully fledged data base system, including importing and exporting from and to other data base systems and/or “Reference Managers”, plus the use of popular “Reference Managers” that can help authors in citing works and generating bibliographies. Therefore, our system significantly contributes to preserving the considerable investment any individual researcher may be making while reading, annotating, referencing, and contributing to the scientific literature in a systematic manner. This is key to any curation of that investment, which risks becoming lost entirely, which we suspect is typically the case. It is also quite remarkable how many Library and Information Science based software and tools, in particular the most popular ones, fall short of addressing all these issues. This curation gap can only be narrowed if toolmakers understand the actual situation in which today’s researchers have to and are working with the scientific literature.

5. Theory and Models

Scientific theories and their development is beyond the scope of this paper. Yet, a few considerations are in place. First scientific publications capture to a large extent theories, including their context, while preserving all subtleties and face the least risk in terms of curation as discussed in Section 4.2. However, this is not the case for the ever growing branch of science in which models, notably complex simulation models, are used to capture and encapsulate theoretical understanding. They require particular attention not to undermine theory development if the dependency of the latter is likely to grow while the former, the simulation models, that typically depend on quite sophisticated computing technologies, are particularly threatened by software erosion. The arising curation challenges are also to be met in this area.

It is important to distinguish between theoretical models, called base models by some, e.g., [54], which represent a scientific theory trying to explain and/or describe a phenomenon from the observable world paraphrased as “Reality” (Figure 7, brown oval) and simplified, parametrized, or lumped models, e.g., [55–58], the simulation models (Figure 7), that typically aggregate many aspects of the real system. While the former models are usually published in scholarly manuscripts, simulation models are rarely publicly accessible in full.

The models and the data needed to run, analyze, and validate the former are obviously intricately interrelated (see also Figure 7). Yet, our experience showed that if model data are separated from the mathematical structures contained in a model, several advantages arise. One is that the same data can elegantly be used for several model variants, e.g., useful when testing alternative theoretical hypotheses each encapsulated in a different model variant with alternative mathematical equations. Another advantage is that different sets of parameters can be applied to the same mathematical model, e.g., useful during ensemble simulations to mimic stochasticity, a sensitivity analysis or when a model, say a vegetation model, is moved among continents.

To this end, Data Frames were developed by the Terrestrial Systems Ecology Group within a subproject of the project RAMSES (Research Aids for Modeling and Simulation of Environmental Systems, [58], www.sysecol.ethz.ch/ramses/), based on Wymore and Zeigler’s systems theory [56,57]. These are the same Data Frames as mentioned earlier (Section 3) in the context of storing meta data together with original data. Data Frames are mere text files that can be freely

formatted employing a specific LL(1) syntax, e.g., [52], the DTF syntax (see Appendix A.3, Figure A7), and are created and edited by whatever tool the researcher prefers. Data Frames can comprise large and complex data sets, not only for model data, but also to hold large observations (cf. Section 3). This is possible since they can encompass any number of files through file references (see Appendix A.3, Figure A7) hereby constituting an entire data base, yet offering efficient retrieval thanks to filtering and other acceleration techniques.

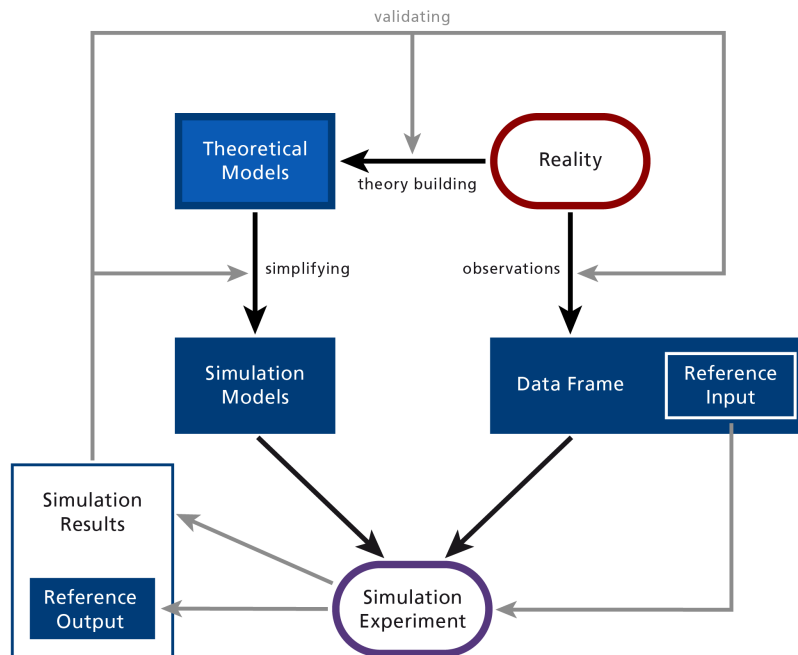


Figure 7. Research cycle of theory building involving model development, simulation experiments, and data analysis by stepwise refinement. The blue boxes denote entities—comprising data and models—that need to be archived for the purpose of meaningful curation, e.g., reuse of models. If archived properly, models can recompute simulation results that were not archived, e.g., due to size. To test a model’s reactivation the archive should also hold the data needed for special simulation experiments that attempt to reproduce the also archived reference output, hereby ensuring reliable model reusability (see text).

Separation of mathematical model properties (e.g., Figure 8) from data (e.g., Figure 9) is powerful in many ways. It offers flexibility for the simulationist as stated above, since a free combination of n Data Frames with the same model or m models using the same data frame allow for the definition of many, *i.e.*, $n \times m$ simulation experiments, e.g., [59]. At the same time it facilitates curation, since there is little need to archive $n \times m$ times model plus data, possibly containing a lot of redundancies, but to archive only n Data Frames plus the code from m model definition programs.

Such an approach is illustrated for the here deliberately simple mathematical model for logistic growth, the DESS (Differential Equation Systems Specification) $\dot{x}(t) = ax(t) - bx(t)^2$ (Figure 8).

For any simulation that model needs parameter values (val) including their domain restrictions (min, max) plus units and other data such as initial conditions. The data frame shown in Figure 9 illustrates this for the model parameters of the logistic equation.

Keeping data strictly separated from mathematical models offers not only the aforementioned advantages, e.g., in terms of flexibility for simulation experiments, but also offers considerable advantages in terms of preservation of such efforts. It enables, in particular, the validation of a once archived and retrieved modelling and simulation setup for reuse.

```

VAR
  s      : System;
  grass  : StateVar;
  grassDot : Derivative;
  c1, c2 : Parameter;

PROCEDURE DifferentialEquation;
BEGIN
  grassDot:= c1*grass - c2*grass*grass;
END DifferentialEquation;

```

Figure 8. Example of a simulation model for a simple mathematical model, *i.e.*, the logistic growth equation $\dot{x}(t) = ax(t) - bx(t)^2$, using Data Frames (Figure 9) for its data needs. Such an approach offers the advantages of being in the simulation model very close to the mathematical form of the theoretical model (Figure 7) while simultaneously offering flexibility and facilitating curation.

```

DATAFRAME ModelParameters;
  REMARK = 'For any logistic growth model';
  MODEL = ANY; KEYCOLUMN = Ident;
DATA:
(*-----*)
  Ident  Descr          val  min  max  unit ;
(*-----*)
  r      'Relative growth rate' 0.7  0.0  10.0 'd^-1' ;
  K      'Carrying capacity'  700.0 0.0  1.0E+38 'g/m^2' ;
(*-----*)
END ModelParameters;
(*-----*)

```

Figure 9. Example of a data frame used in the modelling of the the logistic equation (Figure 8) providing the model parameters r and K . This separation of data from model offers flexibility for the simulationist (see text).

If a special reference input, again stored in form of dedicated Data Frames (Reference Input in Figure 7), and the corresponding simulation results capturing the model behavior under this condition are again set aside as the reference output (Figure 7), a test suite for model reuse is created. According to this approach each model should be accompanied (i) with its reference input consisting of Data Frames defining a particular simulation experiment for testing (dark blue 'Reference Input' in Figure 7) and (ii) the reference output, *i.e.*, the simulation results, for example, in a so-called stash file (dark blue 'Reference Output' in Figure 7). If archived together with the code of the simulation model (dark blue 'Simulation Model' in Figure 7) any reactivation of the model can first be tested by regenerating the reference output and comparing it to the expected reference output as contained in the archive. This ensures that a researcher picking up work on that model at a later point in time on a different machine and setup, can perform first such a test before using the model for new research.

This enables the researcher to ensure high information object quality, while archiving only a few simulation results (*cf.* Section 3, Figure 3). In cases where simulation results are very large, *e.g.*, typically the case for the costly results from climate models, this may significantly reduce archive size without sacrificing reproducibility.

There are four cases to be distinguished here. They result from combining cheap *vs.* costly simulation runs due to long simulation times and/or needing super computers and cheap *vs.* costly archiving due to data size. If both the archive size is small and simulation experiments are cheap no challenges exist. Our approach offers, however, great advantages in the two cases where large archive size is a challenge, but the rendering of the simulation setup including its simulation results can be carried out easily, or when costly simulation experiments produce simulation results of a relatively small size. This approach reaches its limits only when both large simulation results accrue and the rerunning of the simulations is very costly, as is, *e.g.*, the case for general circulation models used in climate research. The way in which to strike the balance of how many simulation results enter the archive needs then to be decided on a case to case basis. Yet, our approach using a reference test suite

with artificial experimental conditions focusing on numerical issues and the use of all involved code would nevertheless be of interest for integrity reasons and may therefore matter in the context of high quality curation, another contribution to narrowing the curation gap.

6. General Discussion and Conclusions

Digital data curation calls for appropriate measures in data **and** research management, ideally right from the beginning of data creation, *i.e.*, before all research using that data is carried out. Such measures can only be implemented by researchers themselves who know their data, the dependencies among them and the needed tools, and understand the relationships between data, models, and theories intimately, see e.g., [13,18,37,42].

To do their part researchers need support, which can be given in the form of more structured workflows. This allows them to conduct better research more efficiently while facilitating curation at the same time. We call this a Structured Research Data Management, see also e.g., [29]. This is of particular relevance in the complex hybrid environments we still have today, where part of the work is done fully digitized while other parts, though diminishing, remain more traditional, *i.e.*, “paper based”. Thus, if research as well as curation can benefit from approaches as we propose them here, it appears we would actually have a win-win situation.

Yet the current widening curation gap will not disappear nor narrow due to these insights alone and institutional curators such as data archives and libraries face considerable challenges when they attempt to take over responsibility according to the OAIS model, e.g., [33]. In the following section we discuss some of the reasons that make us expect the curation gap will continue to stay with us and why the proposed approaches described in this paper cannot overcome the curation gap in its entirety.

Poor tools are sometimes used as an excuse for lack of Research Data Management and other measures to create a more structured research practice that would also facilitate curation. Some efforts have been made to address this need, e.g., [60]. Sophisticated tools, if actually available, could of course support and facilitate data management and curation. If software developers make efforts to better understand researchers and make better suited tools available, we expect, similar to others, that the situation will improve, e.g., [42]. However, it is to be feared that this will not suffice, since either economical incentives may be too low for a limited, and highly diverse researcher clientele, and/or that research will advance so rapidly that a full catching up is forever unlikely.

On the other hand it is also important to note that tools, regardless how close to perfection, can never compensate for weak concepts of what is to be achieved and which factors have to be considered in research. As amply demonstrated in this paper, well designed concepts can be implemented even with limited resources and with tools of limited power by combining tools according to comprehensive, overarching, and most of all solid and thought through concepts. The result achievable through clever combinations is then much more than what could be expected from the components alone. However, such concepts are not widespread, or are challenged in their validity for similar reasons as curation is challenged in general, not the least because they are very difficult to make sufficiently attractive in the arms race between hardware and software advances vis-a-vis the fast pace of today’s research progress.

Fortunately, the curation gap can at least be narrowed, as the case study demonstrated. Remarkable success in terms of curation was achieved in particular in major areas where challenges are presently large, yet highly relevant for research in many disciplines, *i.e.*, for (i) data management, (ii) literature management, and (iii) theory advancement in fields using complex models.

The pragmatic approach of making good use of tools and technique currently available as we propose here is likely to stay with us for many reasons, some mentioned above. Another one, probably quite relevant, is that today’s researchers see themselves as confronted with many diverse, partly incompatible research environments to which they have to adapt continuously. Abstract, *i.e.*, general, yet comprehensive and robust concepts can then be applied nevertheless, since they depend little on particular tools and their idiosyncrasies. That is why we argue that for narrowing the curation

gap it is key to have a strong focus on general applicability and to organize work first of all according to general principles rather than specific solutions, e.g., [29].

Our case study shows this strikingly for literature management (Section 4.2), where adequate tools were largely missing, in particular during earlier years. Thanks to the robust and general nature the approach was and could then be extended as new and/or better tools became available. As the literature management in particular also showed, we are still quite far from meeting the actual needs of researchers, and more robust tools are needed to facilitate research practices, given they are designed well, support data exchange flexibly, and are otherwise well suited to today's researcher's situation. Researchers would of course also be helped more already if some of the presented principles, e.g., separating data from models, were transferred from one discipline to another. This does not question our approach, but rather calls for an improvement in terms of further generalizations and further refinement of the concepts emphasized in this paper.

The following lessons can also be learned from our case study: The archive survived well and has flawlessly served its purpose since 1989 till today, hereby surviving several technological changes such as computer platform changes, many operating system and application changes without a glitch and without requiring any extra maintenance due to those changes. The minimalist approach appears therefore to have critically contributed to this success, emphasizing the value of doing the possible while better solutions are just not available or cannot be afforded. In this context it is worth noting that any less minimalist approach in earlier phases of the long-term project preceding the here presented archive has failed and caused critical data losses, still not resolved while attempts are made to salvage as much as possible, e.g., by refitting the defective hardware needed to access those archives.

This experience also somewhat challenges or at least triggers some reservations towards the merit of large data sharing and collaboration initiatives, e.g., TERENO (Terrestrial Environmental Observatories, teodoor.icg.kfa-juelich.de, e.g., [61]) or TERN (Terrestrial Ecosystem Research Network, www.tern.org.au), where it is still to be demonstrated that (i) the coordinating umbrella efforts will and can be maintained on a long-term basis regardless of technological or organizational changes and/or (ii) that individual contributions can preserve their data accessibility fully on a permanent basis. To mention a quite striking example, our experience, having been involved in the similar LEMA (Long-Term Ecosystem Modelling Activity) network of the GCTE project (Global Change and Terrestrial Ecosystems) of IGBP (The International Geosphere-Biosphere Programme: A Study of Global Change, e.g., [62], for URL see [63]), indicates that the aforementioned risks are real, since the LEMA network has now disappeared, e.g., [64], despite having been considered a critical cornerstone of GCTE, e.g., [62]. Finally a lesson to be learned from this and comparable efforts, e.g., [34] is that the researchers' contribution is most critical for success, similar to our initial postulation and to others' findings [18,29,37,42,47].

Finally, another fundamental prerequisite for success in narrowing the curation gap, notably on the longer run, is a clear commitment from the research group leader, including the skills to ensure that all staff members comply by structuring their work according to the agreed concepts.

However, an improved research practice only covers the early phases of the research data lifecycle. Efforts in research practice must be complemented by more institutional responses [26,29,30]. Universities, funders, and journals can play an important role as well, e.g., by setting proper incentives for a research practice improved in terms of curation, e.g., [17,19,25,28]. Notably, more stringent enforcement or extension of scope of existing guidelines, might already trigger some progress, e.g., by applying existing publishing guidelines to supporting material instead of leaving all responsibility with the authors.

Furthermore, universities, libraries, archives, and data centres must continue and strengthen their efforts to meet the curation gap. First by continuing to support and educate researchers. This also requires a thorough understanding of how researchers actually work and the challenges they have to meet. Secondly, data curation services, which certainly come into play once research data

have left the researcher's individual domain, should be improved and strengthened. If curators also do their part here, an integrated approach may result considerably quicker.

Overall, not only would qualities in all involved sectors be increased through a mutual win-win due to considerations as described here, but the movement towards open access to scientific results might also be nurtured. Not the least as a result of minimizing the additional efforts needed to make well organized data openly available. For sure, efforts by both the research and the curation communities, in the directions proposed here will contribute towards narrowing the curation gap, and perhaps even make it disappear eventually. The trend to more digitization may then even help overcome the curation gap instead of contributing to it as is too often the case today.

In the terms of the future outlook of such efforts the issue of digital sustainability is worth mentioning here. The move towards the externalization of the costs will create an imbalance in the long run. The society needs to be aware of the need for digital sustainability where relevant research data is securely stored for generations to come, readily available and curated by professionals, e.g., [15,22]. All this only supports our plea for a pragmatic approach, similar to what 'sheer curation' tries to accomplish.

Digital sustainability not only means the sustainable preservation of data, ensuring the social dimension of sustainability, but encompasses the economic and environmental sustainability as well. Managing and preserving research data more and with a better structure can help advance scientific progress. For example it can help researchers avoid unwittingly repeating research they did not know was conducted before, perhaps because they merely had no access to the details of that research. This in turn economizes funding, saves time and natural resources used to conduct the research, perhaps hereby even contributing to sustainability in general.

Acknowledgments: The authors would like to thank Angela Gastl and Marion Wullschleger (ETH-Bibliothek), as well as two anonymous reviewers, for helpful comments on the manuscript, and Jeremy Wayne Petrus (ZHdK) for technical support on figure creation.

Author Contributions: The authors worked closely together and each author contributed equally to the manuscript. Matthias Töwe brought the viewpoint of libraries and long-term preservation institutions to the author team, Andreas Fischlin provided his experience with the research management system as developed and employed within his Systems Ecology research group, as well as managing the literature during the manuscript creation, and Ana Sesartić was bridging the gap between these two worlds, having worked at both the Systems Ecology research group and now at the ETH-Bibliothek.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Details on Archiving Concept of Terrestrial Systems Ecology Group

The minimalist archiving concept as introduced in Section 2 requires an archive entry description for each archive entry (AEDescr, red, Figure 1). These meta data have to be provided in the form of a separate ASCII text file and have to be added to the archive entry itself just before archiving to ensure that reconstruction of meta data from the archive itself would be possible if the need should arise. Figure A1 illustrates what meta data the archivist, typically the author, has to provide for an archive entry entitled "Easy ModelWorks <= 1.3r2".

To be able to find an archive entry needs it also needs information on the media (AEMedium, blue, Figure 1) on which it is currently stored (Figure A2). This is a separate ASCII text file that is not to be added to the archive entry itself, since it is likely to change due to archive maintenance, e.g., if an archive entry is moved from an old, ageing medium to a new medium. This separation allows the archive content to be kept consistent at all times, regardless of changes in the storage media (Figure 1).

In order to standardize the creation and maintenance of archive entries, instruction manuals guide individual researchers through the creation of an archive entry and its meta data. Other guides instruct users how to search and retrieve archive entries and how to maintain the archive.

```

-----
TITLE.....Easy ModelWorks <= 1.3r2
VERSION.....1.0 - 1.3r2
LIST_OF_PARTS.....see " PartList.Easy ModelWorks <= ..."
                    content summary:
                    :Easy ModelWorks 1.0a.sea
                    :Easy ModelWorks 1.3 (r2).sea
                    :EasyMW 1.0a:
                    :EasyMW 1.3r2:
                    :EasyMW 1.3r3 (now on SED):
AUTHORS.....A. Fischlin
DESCRIPTION.....All development versions of Easy ModelWorks before
                    it has been officially incorporated in the RAMSES release
                    (contains sources and release folders)
HANDLING.....Finder copy, self extracting archives
REMARKS.....a folder "Models" with Easy ModelWorks model definitions
                    is to be found insied the project "DESolver"
SUB_PROJECT_OF.....RAMSES
CROSS_REFERENCES....DM, MW, DESolver
ARCHIVIST.....Frank Thommen
CREATION_DATE.....5.6.1996

```

Figure A1. Example of an archive entry description (depicted in red in Figure 1).

```

-----
MEDIA.....MO disks
LOCATION.....Tresor CHN C-stock / Schrank CHN D117
NAME.....SE 1 / SE 2
-----

```

Figure A2. Example of the media description for the same archive entry as described in Figure A1. Such media descriptions are depicted in blue in Figure 1.

The workflow for a new archive entry:

- (1) **Discuss the planned entry with the group member responsible for archiving, as well as the group leader as appropriate.** Is the moment right, what should be included, what was or will be archived elsewhere?
- (2) **Prepare archive entry by moving and preparing all elements to be archived to a single folder as well as by sorting and deleting obsolete or redundant files.** This step is crucial as it is where critical files are possibly transformed into file formats with a longer life expectancy (e.g., Word documents → rtf, text, or spreadsheets → SYLK files). Despite the principle of avoiding redundancy, the latter redundancy is intentional as it should help to maximize the archive's long-term use. To this end additional copies are to be saved from critical documents that contain less information than their original, e.g., by force-saving a Word document in form of a plain text file hereby losing all formatting information. Thus a scientific article, if, e.g., written with Word, would be saved into the archive entry three times: (i) binary Word file (doc/docx), (ii) rtf (text file with formatting information), (iii) plain text file (text file without any formatting information). Finally to ensure reuse, e.g., for the purpose of an erratum writing or a continuation of the research, all original master files used in the preparation of figures or tables, e.g., statistical procedures such as R scripts, are included in the archive entry when archiving a scientific article, despite the publication of the latter. E.g., if using \LaTeX , all \LaTeX -files needed as input for the full typesetting enter the archive in their original form in addition to the published PDF. The archive entry is only considered ready when procedures such as typesetting can be performed using the files in the archive. Common parts, e.g., modelling and simulation software, shared by many researchers can be left out, but such dependencies need to be discussed within the research team to ensure their parallel archiving is warranted.

- (3) **Create meta data, *i.e.*, the archive entry description**, for the archive entry (Figure A1). Each entry has to contain information regarding project title, version, list of parts, authors, description of the content, handling (which comprises also which software and/or hardware is necessary to read the data), general remarks, whether the project is part of a sub-project and has any cross-references, the archivist and creation date.
- (4) **Add the archive entry description (Figure A1) to the archive entry itself**. Optionally—particularly useful in case of large and complex archive entries—as the very last step affecting the archives content, some file listing tools can be used to scan the entire archive entry and add a detailed and exhaustive file list part to the archive entry, again in form of a text file.
- (5) **Decide on the type and number of storage media required**, e.g., magneto-optical (MO) disks, CDs, or DVDs, typically determined by the size of the archive entry. Use disks, e.g., MO disks, previously used if not yet full and the entry fits on it. Use multiple disks, e.g., DVDs, if the archive entry is too large to fit on a single disk (n:n relationship).
- (6) **Write the media description on the archive entry** to the separate small text file. These meta data describe the storage media, location and name of the media (Figure A2).
- (7) **Save the archive entry to the archives twice**. Once the above steps are completed to the satisfaction of the peer responsible for archiving, save the ready archive entry by copying twice the entire file system branch to the target storage media. Both copies must be identical and made to the same media, e.g., MO disk, CD, or DVD.
- (8) **Append the archive entry description to the global index** stored on a central file server (Figure 1).
- (9) **Append the archive entry's media description to the global index** stored on a central file server. The global index now contains all meta data allowing access to any archive entry (Figure 1).
- (10) Optionally **store the entry's media description to a separate, also centrally stored media index** tabulating all available archive media. That index is not critical (not shown in Figure 1) and only informs the archivist when periodic maintenance is due, *i.e.*, which parts of the archive should be copied to new media as media ages and which media descriptions need updating.
- (11) Finally, **store the archive media**, e.g., two MO-disks, two CDs, or two DVDs, in two separate locations, *i.e.*, in a safe and an archival cabinet in another building.

The security and long-term preservation of the archive was ensured by several measures: First by using a disaster proof (fire, water, radioactivity, electromagnetic pulse) safe. Secondly by storing a second copy of the entire archive at another location. Thirdly preventive periodic archive maintenance, using the separate storage media index, ensured readability by moving archive entries to new media as their media ageing and/or platform changes required (following inspection/copy intervals are observed: five years for DVDs and floppy discs, ten years for CDs and MO disks). Fourthly, any loss of the meta data kept outside the archive, caused for instance by a fire burning the global index, does not risk the archive, since it is self-contained, *i.e.*, all meta data, including the storage media information, could be reconstructed from the archive itself, albeit the latter would be rather laborious.

Access was restricted to regular and previous employees, while personal data requires particular additional authorization, e.g., from the leader or its deputy, to protect the privacy of possibly involved personal data. This approach comprises Private and Shared, but not Public Domain as shown in Figure 2.

The retrieval workflow was comprised of three steps: (i) Searching for an entry within the global index, (ii) ensuring that one has the access rights, and (iii) copying the needed archive entry from the archive media to a presently used computer system. Archive users were asked to clarify the purpose of the retrieval, whether the data were retrieved for read-only purposes or for continuing research.

In the former case users had to delete the retrieved copy after use to avoid accidental re-archiving. In the case of reuse for new research, users had to make sure that the data are only archived again if actually modified.

Appendix A.2. Reference Management

This describes one concrete implementation of the concept for literature management presented in Section 4.2. Some fundamental design principles need to be observed to implement the proposed literature management system. Among those are the so-called reference types (Figure A3).

| | |
|-------------------------------|---|
| <i>Journal Article</i> | Scientific journal article |
| <i>Electronic Article</i> | Scientific journal article published in electronic form requiring to cite it via doi |
| <i>Newspaper Article</i> | Any non-scientific magazine or newspaper article |
| <i>Book</i> | A book that has an ordinary publisher and all has been written by the same authors and has NO EDITORS! |
| <i>In Book</i> | An individual chapter or section out of a book. NOTE: The authors are the same as for the entire book! Don't confound with <i>In Edited Book</i> ! |
| <i>Edited Articles</i> | A published collection of <i>Journal Articles</i> having editors typically bound as a book that represents a special issue of the scientific journal |
| <i>Edited Book</i> | Like a <i>Book</i> , but has editors |
| <i>In Edited Book</i> | An individual contribution to an <i>Edited Book</i> written by specific authors (who may or may not be the same persons as the editors) |
| <i>Thesis</i> | A work such as a master or PhD thesis as conducted at an university |
| <i>Map</i> | A work that is a map |
| <i>Report</i> | A work that has not been published by an institution (not a publisher) and has no editors |
| <i>Technical Report</i> | Same as the <i>Report</i> , but of a more technical nature, e.g. manuals or detailed model descriptions |
| <i>Edited Report</i> | Like a <i>Report</i> but has editors |
| <i>Proceedings</i> | Proceedings of a conference that can not be cited as an <i>Edited Book</i> |
| <i>In Proceedings</i> | A work that is a contribution contained in <i>Proceedings</i> |
| <i>Booklet</i> | A work that is printed and bound, but without a named publisher or sponsoring institution. |
| <i>Web page</i> | |
| <i>Unpublished</i> | Unpublished work having an author and title |
| <i>Personal Communication</i> | Findings learned about through personal communications (oral testimony) |
| <i>Miscellaneous</i> | A work that matches none of the above |

Figure A3. Types of references used in the proposed literature management system consisting of *LitCentral* and its personal offshoots *LitMYx* (cf. Figure 5) implemented using *FileMaker* (Bib $\text{T}_{\text{E}}\text{X}$: entry types; *EndNote*: Reference Types; *Mendeley*: Type, etc.). The second column offers some explanations about the reference type and its use. The colors together with the sequence define priorities (top green highest, bottom red lowest) by which reference types ought to be used in cases where multiple possibilities exist.

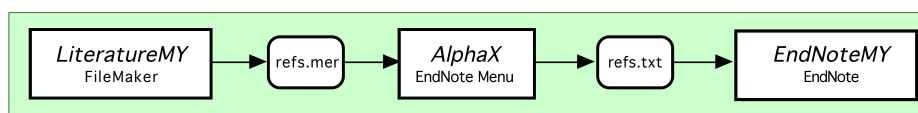
The references need to be comprehensively defined, mappable among all supported tools, notably specific “Reference Managers” such as Bib $\text{T}_{\text{E}}\text{X}$ based *BibDesk* or *EndNote* or *Mendeley*. A task far from trivial. The set shown in Figure A3 has been in successful use many years and satisfies all requirements. However, it is mostly targeted for natural scientists and may not adequately support social scientists and humanities scholars. In particular, the set is likely to require some extension to serve legal scholars equally well. However, and more importantly, Figure A3 proposes reference types that are used commonly, yet surprisingly lack in some of the more popular reference managers (e.g., *Mendeley*). Among those are the reference types *EditedBook*, *InEditedBook*, or *InProceedings*,

which, if lacking and consequently mix, e.g., *Book* with *EditedBook*, prevent the generation of a proper and unambiguous list of references that, e.g., clearly distinguishes between editors and authors.

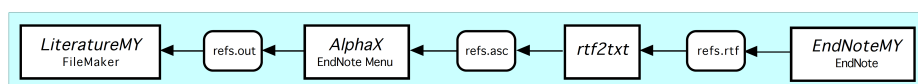
Figure A3 also recommends how to use these reference types, notably when a researcher may be in doubt about what to use or otherwise requires clear guidance. Ex.: A special issue of a scientific journal containing the proceedings of a conference could be cited using reference type *Edited Articles*, *Edited Book*, or *Proceedings*. An individual paper published in such proceedings could then be cited using reference type *Journal Article*, *In Edited Book*, or *In Proceedings*. This table recommends favouring *Edited Articles* over *Edited Book* or *Proceedings* and to favour *Journal Article* over *In Edited Book* or over *In Proceedings*. These priorities are expected to help readers and librarians locate the cited works and help avoid the common mistake of using the physical appearance of a work, e.g., being a book, as a basis for deciding which reference type to use.

The proposed concept requires the transfer of records freely among the involved tools such as the *FileMaker* data base of pivotal importance within the domain of every researcher, project specific *EndNote* bibliographies, e.g., when writing papers with *Microsoft Word*, and project specific Bib_{TEX} files, e.g., when writing papers with L^AT_EX. The central data base storing all records owned by a user is only the *FileMaker* data base (green records, Figure 5). Any project specific storage of read-only copies in *EndNote* bibliographies or Bib_{TEX}/*BibDesk* files (orange records, Figure 5) contains an excerpted subset of the records. Any update of green records requires a retransfer from *FileMaker* to the project specific files *EndNote* bibliographies or Bib_{TEX}/*BibDesk* files, overwriting any orange records possibly stored there (cf. Figures A4 and A5).

1 From *FileMaker* to *EndNote* (→):



2 From *EndNote* to *FileMaker* (←):





-  Temporarily needed text file
-  Application or tool

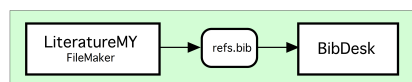
Figure A4. Transfer records between *Filemaker* data base *LiteratureMY* and a project specific *EndNote* bibliography *EndNoteMY* (see also Figure 5). Transfers are possible in both directions, while the transfer from *EndNote* to *FileMaker* is only used during the entering of new records. The use of any intermediate tools such as the tcl based open source Text Editor *AlphaX* are fully automated and require no user intervention.

Transfer concepts also enable the use of multiple devices (Figure 5). Whether researcher x uses more than one device, e.g., a laptop plus a tablet, is then completely x 's decision. Notably synchronizing of records among multiple devices owned by the same person x is also only the personal responsibility of researcher x . Here (Figure 5) we assume conceptually that when the same person uses multiple devices this means basically only a duplication of domain A on several devices, where it is left to the individual researcher's responsibility to make sure she is working only with the master copy at any given point in time.

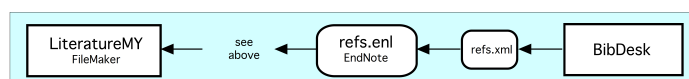
Fortunately currently available tools already provide options, e.g., via the so-called cloud, to synchronize bibliographic meta data and full texts among personally owned devices. Yet, in cases of technical shortcomings or unreliability of automatic syncing techniques, technically simple manual synchronization from a dedicated master device to another, e.g., using a classical syncing tool at the file level (such tools proven to be useful to that end are also offered as "Free Software"

www.sysecol.ethz.ch/people/afischli/software by one of the authors), can also help to overcome today's numerous shortcomings of the cloud services currently on the market. Again, our proposed system offers full flexibility here.

1 From *FileMaker* to *BibDesk* (→) (FM Script *Export to BibDesk*):



2 From *BibDesk* to *FileMaker* (←) (FM Script *Import from BibDesk*):



- Temporarily needed text file
- Application or tool

Figure A5. Transfer records between *Filemaker* data base *LiteratureMY* and a project specific BibTeX file *BibDesk* (see also 5; *BibDesk* is a versatile, powerful open source manager for BibTeX files). Transfers are possible in both directions, while the transfer from *BibDesk* to *FileMaker* is only used during the entering of new records. The use of any intermediate tools such as *EndNote* or the tcl based open source Text Editor *AlphaX* are fully automated and require no user intervention (cf. Figure A4).

| <i>EndNote X1..X7</i> | <i>LiteratureMY</i> | <i>BibDesk (BibTeX)</i> |
|---|---------------------|----------------------------------|
| Aggregated Database | - | - |
| Ancient Text | - | - |
| Artwork | - | - |
| Audiovisual Material | - | - |
| Bill | - | - |
| Blog | - | - |
| Book | Book | book |
| Book Section | In Edited Book | incollection |
| Case | - | - |
| Catalog | - | - |
| Chart or Table | - | - |
| Classical Work | - | - |
| Tech Report (Computer Program) | Tech Report | manual |
| In Proceedings (Conference Paper) | In Proceedings | inproceedings, conference |
| Proceedings (Conference Proceedings) | Proceedings | proceedings |
| Dictionary | - | glossdef |
| Edited Book (Edited Book) | Edited Book | book |
| Electronic Article (Electronic Article) | Electronic Article | electronic |
| Electronic Book | - | - |
| Electronic Book SEction | - | - |
| Encyclopedia | - | - |
| Equation | - | - |
| Figure | - | - |
| Film or Broadcast | - | - |
| Generic | Miscellaneous | misc |
| Generic | Booklet | booklet |
| Government Document | - | - |
| Grant | - | - |
| Hearing | - | - |
| Interview | - | - |
| Journal Article (Journal Article) | Journal Article | article |
| Legal Rule or Regulation | - | - |
| Magazine Article | - | - |
| Manuscript | - | - |
| Map | Map | misc |
| Newspaper Article | Newspaper Article | article |
| Online Database | - | url |
| Online Multimedia | - | url |
| Booklet (Pamphlet) | Booklet | booklet |
| Custom (Patent) | Custom | misc |
| Personal Communication | - | - |
| Podcast | - | - |
| Press Release | - | - |
| Report | Report | techreport |
| Serial | - | periodical |

Figure A6. Mapping of reference types between the involved systems, i.e., *Filemaker* data base *LiteratureMY*, *EndNote*, and *BibTeX*.

Transferring records (Figures A4 and A5) of course requires the mapping of reference types (Figure A3) from one storage place to another. Figure A6 describes the mapping between the *FileMaker*, *EndNote*, and *BibTeX*.

Appendix A.3. Modelling and Simulation

Data Frames, e.g., [59] were developed by the Terrestrial Systems Ecology Group in the context of the ISIS (Integrative Systems Implementation Software, www.sysecol.ethz.ch/ramses/layer/ISIS) in analogy to the system's theoretical concept of Experimental Frames, e.g., [54]. Data Frames allow data to be stored in free format on simple text files and are written in a specific DTF syntax (Figure A7). The DTF syntax is LL(1), e.g., with concise explanations [52] for easier and efficient parsing and allows meta data to be injected anywhere, making the data storage very intuitive to understand (e.g., Figures 4 and 9).

```

DataFrameFile      = (FileReference | DataFrame)
                   {FileReference | DataFrame}.
FileReference      = "FILE" "=" fileName ";" [FilterSpecif].
fileName           = STRING.
FilterSpecif      = "USE" "IF" "FILTER" "=" filterVal ";".
filterVal          = LONGINT.
DataFrame          = "DATAFRAME" dataFrameIdent ";"
                   [DataFrameParamList]
                   "DATA" ":" Table
                   "END" dataFrameIdent ";".
dataFrameIdent     = IDENTIFIER.
DataFrameParamList = { [FilterSpecif] | [ParentOrModelSpecif] |
                      [RemarkSpecif] | [KeyColumnSpecif] }.
ParentOrModelSpecif = ("PARENT"|"MODEL") "=" parentOrModelID ";".
parentOrModelID    = IDENTIFIER | "ANY" | "ALL".
RemarkSpecif       = "REMARK" "=" STRING ";".
KeyColumnSpecif    = "KEYCOLUMN" "=" keyColumnID ";".
keyColumnID        = IDENTIFIER.
Table              = TableHeader TableLine {";" TableLine}.
TableHeader        = IDENTIFIER {IDENTIFIER} ";".
TableLine          = TableEle {TableEle}.
TableEle           = (INTEGER | LONGINT | REAL | LONGREAL |
                     IDENTIFIER | STRING | BOOLEAN).

```

Syntax of the elementary data types is (regular expression notation):

```

INTEGER           = [+]?[0-9]+
LONGINT           = [+]?[0-9]+ "D"
REAL              = [+]?[0-9]+ "." [0-9]+ (("E"|"e")[+]?[0-9]+)?
LONGREAL         = [+]?[0-9]+ "." [0-9]+ (("D"|"d")[+]?[0-9]+)?
IDENTIFIER        = [a-zA-Z] [_a-zA-Z0-9]*
STRING            = ('.*)"('."*)
BOOLEAN          = "TRUE" | "FALSE"

```

Comments start with "("*" and close with "*"*)" and may be nested.

Figure A7. Extended Backus-Naur definition of the DTF syntax used in ISIS (Integrative Systems Implementation Software) for storing data in the form of so-called Data Frames. The latter were developed by the Terrestrial Systems Ecology Group. The DTF syntax is LL(1) [52] and enables flexible, e.g., multi file data storage and efficient random access at run time, e.g., when the data are accessed by model definition programs, e.g., [58] (e.g., Figures 4, 8 and 9).

For efficient retrieval and modular storage data can be stored in a distributed manner using conditional file references (see *FileReference* construct, e.g., *FILE* = "~/DataBases/TemperatureObs/CH1887-1921.DTF; USE IF FILTER = 1880"). A data frame must be stored in its entirety within a single file, while a file may contain *n* Data Frames (e.g., Figure 4). If during a retrieval multiplied definitions of data frame with the same identifier are encountered, only the data from the last read are actually accessible, while circular file references abort the scanning of the involved branch as the circle closes.

Retrieval of the information stored by Data Frames is done via so-called value definitions. The Data Frames allow random access to data via an identifier, decoupling the storage sequence from

the use sequence. This decoupling also allows access to data at run-time from within model definition programs [58] and provides a very flexible and theoretically sound simulation environment. Using the data frame from Figure 4 the identifier denoting the scientific name of the tree species European silver fir would be “ForClim.Aalb.Scientific_name” (scalar value definition) while the vector of bucket sizes from all sites available would be denoted by the identifier “ForClim.Bucketsize” or the IDs of all sites by “ForClim.SiteID” (vector value definition).

Data Frames know only four elementary data types: real and integer numbers, strings, and identifiers. Numbers can be given in any precision and their actual handling depends on the platform. Numerical problems may arise during conversion to binary formats if an old archive is reused on a different platform than where it was originally archived. If efficiency is of particular concern, Data Frames can of course be processed before use and the data stored temporarily in binary form. Those utilities needed for generating binary repositories would then of course also have to be archived.

References

1. Bunge, M. *Scientific Research I: The Search for System*; Studies in the Foundations Methodology and Philosophy of Science; Springer: Berlin, Germany, 1967.
2. Bunge, M. *Scientific Research II: The Search for Truth*; Studies in the Foundations Methodology and Philosophy of Science; Springer: Berlin, Germany, 1967.
3. Popper, K.R. *Conjectures and Refutations: The Growth of Scientific Knowledge*; Routledge & Kegan Paul: New York, NY, USA, 1963.
4. Popper, K. *Objective Knowledge: An Evolutionary Approach*; Clarendon Press: Oxford, UK, 1972.
5. Pearce-Moses, R. *A Glossary of Archival and Records Terminology*; The Society of American Archivists (SAA): Chicago, IL, USA, 2005.
6. Whyte, A.; Job, D.; Giles, S.; Lawrie, S. Meeting curation challenges in a neuroimaging group. *Int. J. Digit. Curation* **2008**, *3*, 171–181.
7. Baltensweiler, W.; Fischlin, A. The larch bud moth in the Alps. In *Dynamics of Forest Insect Populations: Patterns, Causes, Implications*; Berryman, A.A., Ed.; Plenum Publishing Corporation: New York, NY, USA, 1988; Volume 1, pp. 331–351.
8. Baltensweiler, W.; Fischlin, A. On methods of analyzing ecosystems: Lessons from the analysis of forest-insect systems. *Ecol. Stud.* **1987**, *61*, 401–415.
9. Fischlin, A.; Baltensweiler, W. Systems analysis of the larch bud moth system. Part I: The larch-larch bud moth relationship. *Mitt. Schweiz. Ent. Ges.* **1979**, *52*, 273–289.
10. Ruchti, J.; Fischlin, A.; Strässler, E. *DDLDML. Hilfsprogramm für PASCAL Programmierer zur Definition und Verwaltung von INFOSYS-Datenfiles (MANUAL)*; Institut für Phytomedizin ETHZ (vormals Entomologisches Institut ETHZ): Zürich, Switzerland, 1978.
11. Consultative Committee for Space Data Systems and Secretariat (CCSDS). Reference Model for an Open Archival Information System (OAIS)—Recommended Practice; CCSDS: Washington, DC, USA, 2012.
12. Andre, P.Q.C.; Besser, H.; Elkington, N.; Garrett, J.; Gladney, H.; Hedstrom, M.; Hirtle, P.B.; Hunter, K.; Kelly, R.; Kresh, D.; et al. In *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*; The Commission on Preservation and Access and The Research Libraries Group (RLG): New York, NY, USA, 1996.
13. Knight, G.; Pennock, M. Data without meaning: Establishing the significant properties of digital research. *Int. J. Digit. Curation* **2009**, *4*, 159–174.
14. Hsu, L.; Martin, R.L.; McElroy, B.; Litwin-Miller, K.; Kim, W. Data management, sharing, and reuse in experimental geomorphology: Challenges, strategies, and scientific opportunities. *Geomorphology* **2015**, *244*, 180–189.
15. Stafford, N. Science in the digital age. *Nature* **2010**, *467*, S19–S21.
16. Beagrie, N.; Lavoie, B.; Woollard, M. *Keeping Research Data SAFE 2*; JISC: Salisbury, UK, 2010.
17. Beagrie, N.; Chruszcz, J.; Lavoie, B. *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities*; JISC: Salisbury, UK, April 2008.
18. Deridder, J.L. Benign neglect: Developing life rafts for digital content. *Inf. Technol. Libr.* **2011**, *30*, 71–74.

19. Kuipers, T.; van der Hoeven, J. *Insights into Digital Preservation of Research Output in Europe: PARSE-Insight Survey Report D3.6*; PARSE.Insight Project—Permanent Access to the Records of Science in Europe: Didcot, UK, 2010.
20. Waller, M.; Sharpe, R. *Mind the Gap—Assessing Digital Preservation Needs in The UK*; Digital Preservation Coalition (DPC): York, UK, 2006.
21. Nelson, B. Data sharing: Empty archives. *Nature* **2009**, *461*, 160–163.
22. Pryor, G. (Ed.) *Managing Research Data*; Facet Publishing: London, UK, 2012.
23. Ayris, P.; Davies, R.; McLeod, R.; Miao, R.; Shenton, H.; Wheatley, P. *The LIFE2 Final Project Report; Final Report (22/08/08)*; JISC, British Library, University College London, LIBER: London, UK, 2008.
24. Burton, A.; Groenewegen, D.; Love, C.; Treloar, A.; Wilkinson, R. Making research data available in Australia. *IEEE Intell. Syst.* **2012**, *27*, 40–43.
25. Davidson, J.; Jones, S.; Molloy, L.; Kejser, U.B. Emerging good practice in managing research data and research information within UK universities. *Procedia Comput. Sci.* **2014**, *33*, 215–222.
26. Kuipers, T.; van der Hoeven, J. *Insights into Digital Preservation of Research Output in Europe: PARSE-Insight Survey Report; Insight Report D3.4*; PARSE.Insight Project—Permanent Access to the Records of Science in Europe: Didcot, UK, 2009.
27. Neuroth, H.; Strathmann, S.; Oßwald, A.; Scheffel, R.; Klump, J.; Ludwig, J. (Eds.) *Digital Curation of Research Data—Experiences of a Baseline Study in Germany*; Verlag Werner Hülsbusch: Glückstadt, Germany, 2013.
28. OpenAIRE. *OpenAIRE Horizon2020 Factsheets: Open Research Data Pilot in Horizon 2020*, 2015. Available online: www.openaire.eu/or-data-pilot-factsheet (accessed on 3 June 2016).
29. Pryor, G.; Jones, S.; Collins, E.; Whyte, A. (Eds.) *Delivering Research Data Management Services: Fundamentals of Good Practice*; Facet Publishing: London, UK, 2014.
30. Tenopir, C.; Birch, B.; Allard, S. *Academic Libraries and Research Data Services: Current Practices and Plans for the Future*; Association of College and Research Libraries (ACRL): Chicago, IL, USA, 2012.
31. Tenopir, C.; Sandusky, R.J.; Allard, S.; Birch, B. Research data management services in academic research libraries and perceptions of librarians. *Libr. Inf. Sci. Res.* **2014**, *36*, 84–90.
32. Palaiologk, A.S.; Economides, A.A.; Tjalsma, H.D.; Sesink, L.B. An activity-based costing model for long-term preservation and dissemination of digital research data: The case of DANS. *Int. J. Digit. Libr.* **2012**, *12*, 195–214.
33. Kejser, U.B.; Nielsen, A.B.; Thirifays, A. Cost model for digital preservation: cost of digital migration. *Int. J. Digit. Curation* **2011**, *6*, 255–267.
34. Carlson, D. A lesson in sharing. *Nature* **2011**, *469*, 293–293.
35. Björk, B.C.; Hedlund, T. A formalised model of the scientific publication process. *Online Inf. Rev.* **2004**, *28*, 8–21.
36. Thomas, A.; Campbell, L.M.; Barker, P.; Hawksey, M., Eds. *Into the Wild: Technology for Open Educational Resources*; University of Bolton: Bolton, UK, 2012.
37. Van den Eynden, V.; Bishop, L. *Sowing the Seed: Incentives and Motivations for Sharing Research Data, a Researcher's Perspective*; UK Data Archive, University of Essex: Essex, UK, 2014.
38. Chu, H. Research methods in library and information science: A content analysis. *Libr. Inf. Sci. Res.* **2015**, *37*, 36–41.
39. Tammaro, A.M.; Casarosa, V. Research data management in the curriculum: An interdisciplinary approach. *Procedia Comput. Sci.* **2014**, *38*, 138–142.
40. Jones, S. *How to Develop a Data Management and Sharing Plan*; DCC How-to Guides; Digital Curation Centre (DCC): Edinburgh, Scotland, 2011.
41. Verbaan, E.; Cox, A.M. Occupational sub-cultures, jurisdictional struggle and third space: Theorising professional service responses to research data management. *J. Acad. Libr.* **2014**, *40*, 211–219.
42. Goodman, A.; Pepe, A.; Blocker, A.W.; Borgman, C.L.; Cranmer, K.; Crosas, M.; Di Stefano, R.; Gil, Y.; Groth, P.; Hedstrom, M.; *et al.* Ten simple rules for the care and feeding of scientific data. *PLoS Comput. Biol.* **2014**, *10*, e1003542.
43. Foster, N.F.; Gibbons, S. Understanding faculty to improve content recruitment for institutional repositories. *D-Lib Mag.* **2005**, *11*, 1–12.
44. Tenopir, C.; King, D.W. The use and value of scientific journals: past, present and future. *Serials* **2001**, *14*, 113–120.

45. Tenopir, C.; King, D.W.; Spencer, J.; Wu, L. Variations in article seeking and reading patterns of academics: What makes a difference? *Libr. Inf. Sci. Res.* **2009**, *31*, 139–148.
46. Morrow, T.; Beagrie, N.; Jones, M.; Chruszcz, J. *A Comparative Study of E-Journal Archiving Solutions*; Joint Information Systems Committee (JISC): Bristol, UK, 2008.
47. Sutter, R.D.; Wainscott, S.B.; Boetsch, J.R.; Palmer, C.J.; Rugg, D.J. Practical guidance for integrating data management into long-term ecological monitoring projects. *Wildl. Soc. Bull.* **2015**, *39*, 451–463.
48. Waldrop, M.M. The origins of personal computing. *Sci. Am.* **2001**, *285*, 84–91.
49. Treloar, A.; Harboe-Ree, C. Data management and the curation continuum: How the Monash experience is informing repository relationships. In Proceedings of 14th Victorian Association for Library Automation, Conference and Exhibition, Melbourne, VIC, Australia, 5–7 February 2008.
50. Treloar, A. Private Research, Shared Research, Publication, and the Boundary Transitions, Version 1.4.3, 19 Mar 2012. Available online: andrew.treloar.net (accessed on 8 June 2016).
51. CCSDS. *Space Data and Information Transfer Systems—Open Archival Information System (OAIS)—Reference Model*; ISO 14721:2012; Consultative Committee for Space Data Systems and International Organization for Standardization (ISO): Geneva, Switzerland, 2012.
52. Bongulielmi, A.P.; Cellier, F.E. On the usefulness of deterministic grammars for simulation languages. *ACM SIGSIM Simul. Digest* **1984**, *15*, 14–36.
53. Zumstein, P.; Stöhr, M. Zur Nachnutzung von bibliographischen Katalog- und Normdaten für die persönliche Literaturverwaltung und Wissensorganisation. *ABI Tech.* **2015**, *35*, 210–221.
54. Zeigler, B.P. The five elements. In *Theory of Modelling and Simulation*, 1st ed.; Robert E. Krieger Publishing Company Inc.: Malabar, FL, USA, 1976; pp. 27–49.
55. Zeigler, B.P. Multilevel multiformalism modeling: An ecosystem example. In *Theoretical Systems Ecology*; Halfon, E., Ed.; Academic Press: New York, NY, USA, 1979; pp. 17–54.
56. Zeigler, B.P. *Theory of Modelling and Simulation*, 1st ed.; Robert E. Krieger Publishing Company Inc.: Malabar, FL, USA, 1976.
57. Wymore, A.W. Theory of systems. In *Handbook of Software Engineering*; Vick, C.R., Ramamoorthy, C.V., Eds.; Van Nostrand Reinhold Company: New York, NY, USA, 1984; pp. 119–133.
58. Fischlin, A. Interactive modeling and simulation of environmental systems on workstations. In *Analysis of Dynamic Systems in Medicine, Biology, and Ecology*; Möller, D.P.F., Richter, O., Eds.; Springer: Berlin, Germany, 1991; Volume 275, pp. 131–145.
59. Nemecek, T. *The Role of Aphid Behavior in the Epidemiology of Potato Virus Y: A Simulation Study*; Diss. ETH No. 10086; Swiss Federal Institute of Technology: Zürich, Switzerland, 1993.
60. Androulakis, S.; Buckle, A.M.; Atkinson, I.; Groenewegen, D.; Nicholas, N.; Treloar, A.; Beitz, A. ARCHER—E-Research tools for research data management. *Int. J. Digit. Curation* **2009**, *4*, 22–33.
61. Kunkel, R.; Sorg, J.; Kolditz, O.; Rink, K.; Klump, J.; Gasche, R.; Neidl, F. TEODOOR—A spatial data infrastructure for terrestrial observation data. In Proceedings of the 2013 IEEE 10th International Conference on Networking, Sensing and Control (ICNSC), Evry, France, 10–12 April 2013; pp. 242–245.
62. Steffen, W.L.; Walker, B.H.; Ingram, J.S.I. *Global Change and Terrestrial Ecosystems: The Operational Plan*; Global Change Report No. 21; International Geosphere-Biosphere Program (IGBP): Stockholm, Sweden, 1992.
63. GCTE - Global Change and Terrestrial Ecosystems. Available online: www.igbp.net/researchprojects/igbpcoreprojectsphaseone/globalchangeandterrestrialecosystems.4.1b8ae20512db692f2a680009018.html (accessed on 3 June 2016).
64. Seitzinger, S.P.; Gaffney, O.; Brasseur, G.; Broadgate, W.; Ciais, P.; Claussen, M.; Erisman, J.W.; Kiefer, T.; Lancelot, C.; Monks, P.S.; et al. International Geosphere–Biosphere Programme and Earth system science: three decades of co-evolution. *Anthropocene* **2015**, *12*, 3–16.

