

Data sharing in a technological-driven research environment

João Aguiar Castro

Faculty of Engineering. University of Porto, Porto, Portugal (joaoaguiarcastro@gmail.com)

Abstract

Analogous to the technological developments, research data is being produced at a higher rate than ever, making research data management a current scientific issue. In this context researchers are encouraged not only to disclose, but also actively manage their data, including data description, so others can benefit from it. Controlled vocabularies, ontologies and metadata schemas are examples of tools that can support researchers to communicate their work. This paper consists in an overview on how metadata can be a useful asset to improve data sharing between researchers, given the biodiversity domain as an example to illustrate this scenario. Being aware that institutions usually struggle to keep up with the demands of a fast-paced, multi-domain research environment it is arguable that researchers, if provided with tools that best fit their requirements, should play a central role in the research data management workflow.

Subject Headings. Information Science, Biodiversity

Author Keywords. Research Data Management, Ontology, Metadata

1. Introduction

Today digital content is the dominant form by which information is created, shaped and disseminated. Governments, companies, research organizations, libraries and individuals have become highly dependent of digital information. However, there is an awareness that these assets may be at risk with weak guarantees of future access due to the dependence on a technology infrastructure that tends to evolve at a breakneck pace (Li et al. 2013). Simply, everything that is digital is fragile and more likely to fall into disuse, meaning that digital resources are less permanent than their equivalents in paper.

If the emerging digital technologies helped to shape science, arguing that research needs also lead to the development of many technological tools is also true. It is common to use the concept of "e-science" to outline the change to a more computationally driven research environment (Gore 2011). As the research world becomes more digital it imposes new challenges to the longevity and retrieval of scientific information. Yet, the general notion is that, once stored in digital form, information is safeguarded (Smit et al. 2011). Consequently research data is rarely properly stored, and as a result can be lost, sometimes permanently.

Scientists tend to store data with little effort via its preservation, using back-ups on their personal computers, tools like Dropbox, or common storage devices, such as pen-drives. Hence, sooner or later researchers have to deal with data integrity, accuracy and accessibility issues (Borgman 2012).

Furthermore, researchers often do not provide additional contextual information about the data they produce. This type of strategy hardly deals with the amount of data produced, since this approach does not guarantee future data use nor interpretation (Willis et al. 2012). The lack of detailed metadata records raises many concerns in research data management, which are undesirable for data sharing.

In the following section an overview of the research data management topic is provided, particularly addressing researchers' culture towards data sharing. The production of detailed metadata records is needed in this process, and for this reason section 3 discusses the importance of metadata creation. It will also be argued that ontologies can be of great value to support researchers when describing their data. A case study on how the biodiversity community is tackling this issue is also presented here. This paper ends with the discussion between the roles that institutions and researchers must perform while participants in a research environment that is both threatened by the overflow of data, and boosted by emergent technologies.

2. Research Data Management

Research data is a critical aspect in science as it establishes the basis for scientific decision making (Tenopir et al. 2011). Research data not only represent an output of research, but also provide inputs for new hypothesis, that is, data open the way to new scientific knowledge and lead to innovation (National Science Foundation 2012).

More than any other academic output, research data, if well documented, has the potential to impact future research. In the e-science context, the production of research data has reached an unprecedented volume as it became highly instrumented, due to the technological advancement. Therefore, a new paradigm shift was identified to describe "a data intensive scientific discovery" environment (Hey et al. 2009). In fact, literature tends to adopt the term "deluge" to inform about the current large scale data production (Borgman 2012).

Although exponential data production cannot be argued, the deluge harmful effects are occurring only relatively. If "big sciences", like physics and astronomy, are well supported by a set of tools and repositories to handle this intensive data production, in the long tail of science (Heidorn 2008), wherein a large number of small scientific branches are producing a large quantity of heterogeneous data, capable infrastructures to manage the increasing amount of data are needed.

The lack of an integrated framework for managing this type of research data represents a significant barrier not only to those scientists conducting research, but also for those who subsequently will reuse it. From now on, while arguing on research data management issues, these are to be taken of the long tail of science.

Data management needs are not the same for every scientific domain. Research data are different in size and content across scientific domains, and researchers also have different behaviors with respect to data sharing.

Data sharing is an act that can be useful to verify scientific achievements, allowing for instance the verification of results and an extension of the research based on previous results. Sharing includes the deposit and preservation of data, but the main purpose is to enhance access. The broad scientific community benefits from sharing as this encourages new perspectives, allowing different interpretations and approaches that lead to scientific progress, particularly when it takes place in an interdisciplinary background.

One of the arguments that move the scientific community to promote access to data is the opportunity to reuse it. Opportunity that allows, for instance, repeating data collection procedures. Thereby contributing for streamline the research processes. This dynamic can foster the interest of organizations in the potential commercial value of products derived from scientific research, and increase private investment. The development of new services, employability and the production of wealth, are consequences that, in the long term, can arise from research data sharing.

2.1. Research data sharing

Scientific literature on this topic mainly discusses the reasons behind sharing or withholding data between scientists, that is, their perception about making data available. Accordingly, barriers to effective data sharing are rooted in the culture of scientists (Tenopir et al. 2011). The balance between what can be gained or lost from sharing is a concern that always disturbs scientists, and continues to prevail.

Among the main reasons that motivate researchers to retain data is the lack of time or institutional support for data management in the short or long term (Arzberger et al. 2004), commercial interests, confidentiality regarding study participants, future publishing opportunities and retention of exclusivity on the data that led a lot of time to produce (Wicherts et al. 2011). There is also a more perverse motivation to explain this withholding, which is that peers scrutiny of their data can expose errors or produce contradictory conclusions.

Data sharing is beneficial to science, however, one might think that these benefits are less obvious for those who produce or make data available. One of the aspects that researchers value most is the number of times their work is cited, since acquire reputation helps them get funding for further research and may lead to an income augment. It has been shown that researchers that disclose their data are cited more often than those who do not (Piwowar 2007).

Another advantage of data reuse is that it helps to recognize errors and discourages fraud. Furthermore, replication of results serves as a good basis to educate new researchers. In short, data sharing and reuse facilitate science, and in many cases data holds a value that goes way beyond the purpose of its creation in the first place. So, dissemination and unrestricted use of data are essential attributes to enhance scientific innovation.

In this context the ever growing open access movement has been putting pressure on research data sharing. However, a review of the European Commission, revealed a lack of knowledge and understanding of researchers in relation to open access. The latter are not strictly related to the concept, or even if they do, they not demonstrate any knowledge about this issue (European Commission 2010). However, it must be remembered that in some cases researchers might be constrained by restrictions to disclosure, many of them conditioned to embargo periods. Indeed, many scientists consider data sharing important but only few consider access to research data easy (Tenopir et al. 2011).

To shed the emergence on research data management, the National Science Foundation (NSF), an important science funding agency, is recognizing research data as a precious resource. In fact NSF mandates the submission of a research data management plan as a funding requisite, so data “can be routinely deposited in well-documented form, regularly and easily consulted and analyzed by specialist and non-specialists alike, open accessible while suitably protected, and reliably preserved” (National Science Foundation 2012). In fact project funding proposals fail if not addressing research data management properly. The same is valid for projects funded by the European Community, in the Framework Programme for Research and Innovation (Horizon2020) that has established a relation between research data management and its open access policy (H2020 2013).

Despite the identified research data management issues, there is also a lack of guidelines and standardized procedures in creating and storing data, and the metadata produced is scarce. Research data assumes many different properties, and its nature is diverse even in the same domain. Given research data complexity, metadata is necessary for its retrieval,

interpretation, and reuse. However, obtaining comprehensive and accurate metadata is not trivial.

3. Metadata: A key for data sharing

The data deluge is a challenge, both in the quantity and in the multiple formats at which data is created. Storage, by itself, is not a warranty of latter data retrieval and interpretation. If research data does not provide relevant information for others to use it, it won't likely be reused, and eventually will become obsolete. Hence, metadata is of paramount interest in any information system.

Defined by many as "data about data", metadata is useful for resource discovery, contextualization, detailed processing, and ultimately use. In other words metadata can ensure that resources can be curated, accessed, retrieved, and interpreted in the long run. Therefore metadata has become an indispensable component for the description of research data and for scientific communication in general (Greenberg and Feinstein 2013).

Nevertheless, adequate metadata is not easy to obtain. In order to achieved well documented metadata one has to consider the trade-off between the efforts of a time-consuming activity and the benefits that arise from such a commitment.

Research data reuse highly depends on detailed descriptions and in most cases researchers cannot be diverted from the research process itself. Therefore, they can be discouraged to create metadata consistently (Qin and Li 2013). In order to foster the documentation of research data several metadata schemas and vocabularies were developed in recent years. Many resulting from the labor of domain-specific communities, or related to domain-specific data repositories (Willis et al. 2012). Although the benefits of adopting metadata schemas cannot be denied, their proliferation and fragmentation across disciplines can interfere in research data interchange.

Given the sheer number of metadata schemas, and vocabularies, it is expected that the same concept co-occur in many of them. The main issue here is the ambiguity of concept representation, when different terms are used to express the same concept or notion (Qin et al. 2013), compromising metadata effectiveness (Greenberg and Feinstein 2013). This issue can, however, be addressed in the Web-Semantic framework, being ontologies suitable tools to deal with these semantic inconsistencies.

3.1. On the benefits of ontologies

Designed by Tim Berners-Lee, the Semantic Web is defined as "a web of data that can be processed directly and indirectly by machines" (Berners-Lee et al. 2001). The main goal of the Semantic Web is to ease the process to find, combine and share information. This is particularly important since the Internet revolution has deep impact the way scientific information is handled and shared across the web.

By providing an architecture, based on formal semantics, the Semantic Web produce adequate contents for both human and computers to consume (Berners-Lee et al. 2001). In order to succeed, the Semantic Web depends on the development of ontologies that can establish vocabulary alignment for the sake of interoperability (Lassila and Mcguiness 2001).

The term ontology is understood in a variety of ways and has been used in philosophy for many centuries (Smith and Welty 2001). Briefly, in the computer science tradition, an ontology can be defined as a set of concepts, and the relations between them, describing a domain.

A dimension of the convenience of adopting ontologies for data sharing, is that they can assist researchers to design a common vocabulary needed to exchange information in a community of interest (Noy and Mcguiness 2000). Ontology representation concerns expressive, accurate

and non-ambiguous syntaxes, which benefits research data management in this contemporary fast-paced research data production ecology. Therefore, ontologies are an adequate tool to support research data description.

The advantage of using ontologies, while creating metadata records, relies on their ability to make domain-agnostic semantic representations. In a myriad of vocabularies and metadata schemas, ontologies favor interoperability, as they can be viewed as extensions of metadata schemas, matching them to ensure semantic consistency (Qin et al. 2013).

Furthermore, ontologies can be developed asynchronously, as communities can work together, spatial and temporal limitation aside, to incrementally evolve them. Additionally, they are flexible enough to face the data deluge challenges. Essentially, formal ontologies are capable of setting the semantic baseline from which the scientific community can address issues linked to research data description, and eventually escalate data sharing. Along with general purpose structured vocabularies, many communities have developed standardized ontologies that researchers can use to improve metadata (Noy and McGuinness 2000). For instance, given the complexity of biodiversity data, this community has been developed such vocabularies as a way to organize and share data, within and abroad its discipline boundaries.

3.2. Data sharing in the biodiversity domain

Core Semantic Web technologies have brought many metadata opportunities, by founding a web infrastructure where data is linked and the relation between resources is explicitly defined. Likewise, advances in computing and in environmental information systems allowed remarkable discoveries (Jones and Gries, 2010), which lead to a new range of challenges in research data management. The biodiversity domain is a practical example wherein the use of such technologies can be of great significance, since the production of meaningful data is strictly needed.

As the very name suggests, biodiversity data is very diverse. Biodiversity research concerns the ongoing impact of environmental changes, and it relies on the data collected by several multi-disciplinary research groups. Biodiversity data comprises inventories of species names and synonyms, typologies of ecosystem processes, temporal and spatial species distribution, images and sounds, among other facets (Poças et al. 2014; Bach 2012). Academics and worldwide funded research programs are responsible for the data that sustains the biodiversity research. However, if not well documented, this data can be in jeopardy after the conclusion of a given project.

The use of tools and techniques to manage research data can help the biodiversity community to deal with the demanding environmental matters (Jones and Gries, 2010). Accordingly, the biodiversity community is now under pressure, and it is expected a common agreement, and collaboration, on the development of networks and services that favors data sharing (Poças et al. 2014). For instance, the European Union aims to create a spatial data infrastructure that eventually will enable the sharing of environmental information among public sector organizations. Therefore it is essential to adopt standard vocabularies.

For biodiversity researchers data availability is crucial, simply because to obtain field data is an extremely costly activity (Costello and Wieczorek 2014). To expedite data sharing in this domain many databases have emerged, and data storage and disclosure is mandatory in some biodiversity peer-review publications (Jones and Gries 2010).

In order to best achieve metadata consistency in the biodiversity research domain, the production of controlled vocabularies and a general ontology is opportune (Bowker, 2000), for a more proficient approach in capturing the semantics of biodiversity complex datasets,

than just applying natural language metadata (Mena-Garcés et al. 2011). In this case, commitments have already been invested by this community in developing metadata standards, so that understanding and manage a diverse and large amount of data, can be facilitated (Michener 2006).

Depending on the specifics of each biodiversity domain, these standards can go from very particular ones, like the Genomic Contextual Data Markup Language (GCDML), describing “the exact origin and processing of a biological sample, from sampling to sequencing, and subsequent analysis” (Bendix et al. 2012); to more generic ones, such as the Ecological Metadata Language. The Ecological Metadata Language was developed by researchers and information managers, from a wide variety of institutions, with the goal to design a proper standard for describing ecological long-term research data. Albeit its focus on ecology, this standard represent generic terms that serve the interest of a broader community. In its essence, the Ecological Metadata Language was thought to reduce the ambiguity while creating structured metadata and is used in many environmental information management systems (Michener 2006).

Another popular standard in the biodiversity community is the Darwin Core metadata schema. It provides well-defined fields to record taxonomic concepts (Costello and Wieczorek 2014), with the goal to understand a global pattern on biodiversity heterogeneous data sources to improve interoperability (Wieczorek 2012).

Currently, at least a quarter of the existing repositories for biodiversity are applying ontologies, as a flexible method to persistently match concepts attributes and their relations (Bach et al. 2012). An example of a commonly used ontology by the biodiversity community, is the Extensible Observation Ontology, also known as OBOE (Madin et al. 2007). The later targets the description of scientific observation and measurements, particularly highlighting the observational context as information that matters for resource browsing (Madin et al. 2007). Moreover, if combined with the Ecological Metadata Language, OBOE will likely provide the foundation for evolve information systems with formal semantics; enable them to perform tasks like the interpretation of datasets produced at different locations (Mena-Garcés et al. 2011). Thus, enhancing the reuse and sharing of data among the biodiversity researchers.

It is also worth to mention that, despite all the efforts of the biodiversity community in developing a set of tools and vocabularies, still there is plenty to overcome before research data sharing can be fully achieved (Jones and Gries 2010). The next section informs on the convergence needed between researchers and information specialists, so that research data can be documented efficiently and effectively.

4. The role of researchers and curators on data sharing

Academic institutions are ideal backgrounds for providing research data management services. Data curation services were recently point out as one of the top ten trends for academic libraries (Tenopir et al. 2012), and if some are already engaged in research data activities, others are considering to doing so. This data services should ideally be aligned with the current e-science framework, in order to address researchers’ data management requirements. Not only institutions must display the infrastructure to better accommodate these requirements, but they also need to supply services to educate researchers on data management and sharing (Tenopir et al. 2012). Yet, institutions usually lack the infrastructure and struggle to support researcher's requirements.

A possible solution is to have data curators, or other information specialists, to become stakeholders on these data management services. For instance, they can be part of grant

proposal teams as data curation consultants. In big sciences is not unusual to have data scientists, with domain expertise, to perform data related activities. Nevertheless, the same does not apply to small sciences that may fall short in resources to deal with data issues, thus requiring the intervention of academic staff. Data curators assisting researchers on data management can make the difference. Their ability in preservation, security, and accessibility, can improve the value of research data, since they are aware of metadata best-practices, fostering data dissemination.

Information specialists are becoming very participative in this environment. While data curators, and information scientists in general, can make good use of their skills, in the long run their contribution can have less impact, if researchers are not motivated to cope in the overall process. Data curators are not domain-specific experts, or at least not in a wide variety of fields, and have limited practice on research initiatives in general. Therefore it can be a tricky challenge for them to be key stakeholders in data documentation activities.

Data curators can also be overwhelmed by the heterogeneity of scientific disciplines, and its correspondent fragmentation in terms of data practices, that vary widely within the same institution, or even in the same department. To merely depend on data curators to perform data management activities implies a very, and unfeasible, demanding task, being aware that most institutions cannot delegate a data curator for each department of research team.

Despite data curators' doubtful metadata skills, they are only few to provide timely metadata to keep the pace of digital resources, thus creating a bottleneck in the research data workflow. Researchers not only should be empowered of documenting their data, they also must occupy a key position in the development of tools to improve data documentation, such as ontologies. Given their know-how in domain terminology and their regular involvement in research environments, researchers are the most appropriate candidates to produce accurate metadata records. Furthermore, it is good practice to register metadata as soon as possible in the research workflow. Once in the process of creating data it is more likely that researchers hold full knowledge of the research context. On top of that if metadata is not recorded timely it can yield lackluster descriptions (Martinez-Urbe and Macdonald 2009). In this sense, collaboration between researchers and data curators is chief, and both parties should co-exist in the development of vocabularies to support metadata activities.

5. Conclusion and Future Work

Digital technologies and the way they quickly evolve raise as many challenges for research data management, as they bring opportunities for the research community that are still to be fully grasped. The landscape of contemporary science is highly impacted by the large number of research outputs, namely research data. This imposes many challenges, as researchers now have to deal with the correct management of these assets.

Metadata has been described as an important instrument for research data management, particularly by allowing data annotation, in the first place, and its latter retrieval and interpretation. Metadata production does not come without a cost, and technologies had enabled the background for the upsurge of instruments that can be of great value for one to record metadata. The scientific community is increasingly aware of the potential of these instruments, such as ontologies, for enhance data documentation. Hence, communities have gather efforts to deliver controlled vocabularies, and ontologies, for better improve the changes of data reuse, as illustrated by the biodiversity case study.

In this scenario there is a space for researchers and data curators to partner and build vocabularies, taking advantage of their combined skills. In order to data sharing be more likely

to succeed, it is important that researchers pay particular attention to the annotation, documentation and organization of data from the moment they begin to produce it. However, this is a time consumer process and not as motivating as the research process itself, and actions to keep data safe are usually based on ad-hoc practices.

On the other hand, and since research data management is a matter of concern for data curators, these should guarantee the accommodation of research data into repositories, ensuring data correct indexation and classification, so it can be easily retrieved and consequently reused by others. Data curators can act as a connection link between researchers and others communities and working closer with researchers can help the later to better manage and deal with the ever frustrating data deluge. This synergy would certainly contribute to promote a change in the scientific community towards a culture of data sharing, which enables science and the overall society to evolve.

Ongoing research will be informed by this work in the attempt to provide researchers from very specific scientific domain with the metadata models that best fit their data management needs.

References

- Arzberger, P., P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Moorman, P. Uhler, and P. Wouters. 2004. "Promoting access to public research data for scientific, economic, and social development." *Data Science Journal* no. 3:135-152. Accessed May 11, 2016. DOI: [10.2481/dsj.3.135](https://doi.org/10.2481/dsj.3.135).
- Bach, Kerstin, Daniel Schäfer, Neela Enke, Bernhard Seeger, Birgit Gemeinholzer, and Jörg Bendix. 2012. "A comparative evaluation of technical solutions for long-term data repositories in integrative biodiversity research." *Ecological Informatics* no. 11:16-24. Accessed May 11, 2016. DOI: [10.1016/j.ecoinf.2011.11.008](https://doi.org/10.1016/j.ecoinf.2011.11.008).
- Bendix, Jörg, Jens Nieschulze, and William K. Michener. 2012. "Data platforms in integrative biodiversity research." *Ecological Informatics* no. 11:1-4. Accessed May 11, 2016. DOI: [10.1016/j.ecoinf.2012.04.001](https://doi.org/10.1016/j.ecoinf.2012.04.001).
- Berners-Lee, Tim, James Hendler, and Ora Lassila. 2011. "The Semantic Web." *Scientific American*:34 - 43. Accessed May 11, 2016. DOI: [10.1038/scientificamerican0501-34](https://doi.org/10.1038/scientificamerican0501-34).
- Borgman, Christine L. 2011. "The conundrum of sharing research data." *Journal of the American Society for Information Science and Technology*:1-40. Accessed May 11, 2016. DOI: [10.2139/ssrn.1869155](https://doi.org/10.2139/ssrn.1869155).
- Borgman, Christine L., Jillian C. Wallis, and Noel Enyedy. 2007. "Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries." *International Journal on Digital Libraries* no. 7 (1):17-30. Accessed May 11, 2016. DOI: [10.1007/s00799-007-0022-9](https://doi.org/10.1007/s00799-007-0022-9).
- Bowker, Geoffrey C. 2000. "Biodiversity Datadiversity." *Social Studies of Science* no. 30 (5):643-683. Accessed May 11, 2016. DOI: [10.1177/030631200030005001](https://doi.org/10.1177/030631200030005001).
- Costello, Mark J., and John Wieczorek. 2014. "Best practice for biodiversity data management and publication." *Biological Conservation* no. 173:68-73. Accessed May 11, 2016. DOI: [10.1016/j.biocon.2013.10.018](https://doi.org/10.1016/j.biocon.2013.10.018).
- European Comission. 2013. *Multi-beneficiary general model grant agreement - version 1.0*. Accessed May 11, 2016. http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-multi_v1.0_en.pdf.

- European Commission, Directorate-General for Research and Innovation. 2011. *Sharing knowledge: Open access and preservation in Europe, conclusion of a strategic workshop, Brussels, 25-26 november 2010*. Luxembourg: Publications Office of the European Union. Accessed May 11, 2016. DOI: [10.2777/63410](https://doi.org/10.2777/63410).
- Gore, Sally A. 2011. "e-Science and data management resources on the Web." *Medical Reference Services Quarterly* no. 30 (2):167-177. Accessed May 11, 2016. DOI: [10.1080/02763869.2011.562778](https://doi.org/10.1080/02763869.2011.562778).
- Greenberg, Jane, Shea Swauger, and Elena Feinstein. 2013. Metadata capital in a data repository. In *Proceedings of the International Conference on Dublin Core and Metadata Application*. Accessed May 11, 2016. <http://dcpapers.dublincore.org/pubs/article/view/3678>.
- Heidorn, P. Bryan. 2008. "Shedding light on the dark data in the long tail of science." *Library Trends* no. 57 (2):280-299. Accessed May 11, 2016. DOI: [10.1353/lib.0.0036](https://doi.org/10.1353/lib.0.0036).
- Heidorn, P. Bryan. 2011. "The emerging role of libraries in data curation and e-science." *Journal of Library Administration* no. 51 (7-8):662-672. Accessed May 11, 2016. DOI: [10.1080/01930826.2011.601269](https://doi.org/10.1080/01930826.2011.601269).
- Hey, Tony, Stewart Tansley, and eds. Kristin Tolle. 2009. *The fourth paradigm: Data-intensive scientific discovery*: Microsoft Research.
- Jones, Matthew B., and Corinna Gries. 2010. "Advances in environmental information management." *Ecological Informatics* no. 5 (1):1-2. Accessed May 11, 2016. DOI: [10.1016/j.ecoinf.2010.01.001](https://doi.org/10.1016/j.ecoinf.2010.01.001).
- Lassila, Ora, and Deborah McGuinness. 2001. *The role of frame-based representation on the semantic web*. Technical Report KSL-01-02. Stanford, CA: Stanford University.
- Li, Yuan-Fang, Gavin Kennedy, Faith Ngoran, Philip Wu, and Jane Hunter. 2013. "An ontology-centric architecture for extensible scientific data management systems." *Future Generation Computer Systems* no. 29 (2):641-653. Accessed May 11, 2016. DOI: [10.1016/j.future.2011.06.007](https://doi.org/10.1016/j.future.2011.06.007).
- Madin, Joshua, Shawn Bowers, Mark Schildhauer, Sergeui Krivov, Deana Pennington, and Ferdinando Villa. 2007. "An ontology for describing and synthesizing ecological observation data." *Ecological Informatics* no. 2 (3):279-296. Accessed May 11, 2016. DOI: [10.1016/j.ecoinf.2007.05.004](https://doi.org/10.1016/j.ecoinf.2007.05.004).
- Martinez-Urbe, Luis, and Stuart Macdonald. 2009. "User engagement in research data curation." In *Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. Proceedings*, edited by Maristella Agosti, José Borbinha, Sarantos Kapidakis, Christos Papatheodorou and Giannis Tsakonas, 309-314. Berlin, Heidelberg: Springer Berlin Heidelberg. Accessed May 11, 2016. DOI: [10.1007/978-3-642-04346-8_30](https://doi.org/10.1007/978-3-642-04346-8_30).
- Mena-Garcés, Elena, Elena García-Barriocanal, Miguel-Angel Sicilia, and Salvador Sánchez-Alonso. 2011. "Moving from dataset metadata to semantics in ecological research: a case in translating EML to OWL." *Procedia Computer Science* no. 4:1622-1630. Accessed May 11, 2016. DOI: [10.1016/j.procs.2011.04.175](https://doi.org/10.1016/j.procs.2011.04.175).
- Michener, William K. 2006. "Meta-information concepts for ecological data management." *Ecological Informatics* no. 1 (1):3-7. Accessed May 11, 2016. DOI: [10.1016/j.ecoinf.2005.08.004](https://doi.org/10.1016/j.ecoinf.2005.08.004).
- National Science Foundation, Office of Cyberinfrastructure. 2012. *Data Infrastructure Building Blocks (DIBBs)*. Accessed May 11, 2016. <http://www.nsf.gov/pubs/2012/nsf12557/nsf12557.pdf>.

- Noy, Natalya F, and Deborah L McGuinness. 2000. *Ontology development 101: a guide to creating your first ontology*. Accessed May 11, 2016. http://protege.stanford.edu/publications/ontology_development/ontology101.pdf.
- Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. 2007. "Sharing detailed research data is associated with increased citation rate." *PLoS ONE* no. 2 (3):e308. Accessed May 11, 2016. DOI: [10.1371/journal.pone.0000308](https://doi.org/10.1371/journal.pone.0000308).
- Pôças, Isabel, João Gonçalves, Bruno Marcos, Joaquim Alonso, Pedro Castro, and João P. Honrado. 2014. "Evaluating the fitness for use of spatial data sets to promote quality in ecological assessment and monitoring." *International Journal of Geographical Information Science* no. 28 (11):2356-2371. Accessed May 11, 2016. DOI: [10.1080/13658816.2014.924627](https://doi.org/10.1080/13658816.2014.924627).
- Qin, Jian, and Kai Li. 2013. "How portable are the metadata standards for scientific data? a proposal for a metadata infrastructure." In *Proceedings of the 2013 International Conference on Dublin Core and Metadata Applications*, 25-34. Lisbon, Portugal: Dublin Core Metadata Initiative. Accessed May 11, 2016. <http://dcpapers.dublincore.org/pubs/article/viewFile/3670/1893>.
- Qin, Jian, Xiaozhong Liu, and Miao Chen. 2013. "Ontology-enabled metadata schema generator: the design approach." In *Proceedings of the 2013 International Conference on Dublin Core and Metadata Applications*, 1-2. Lisbon, Portugal: Dublin Core Metadata Initiative. Accessed May 11, 2016. <http://dcevents.dublincore.org/IntConf/dc-2013/paper/viewFile/204/207>.
- Smit, Eefke, Jeffrey Van Der Hoeven, and David Giarretta. 2011. "Avoiding a Digital Dark Age for data: why publishers should care about digital preservation." *Learned Publishing* no. 24 (1):35-49. Accessed May 11, 2016. DOI: [10.1087/20110107](https://doi.org/10.1087/20110107).
- Smith, Barry, and Christopher Welty. 2001. "Ontology: towards a new synthesis." In *Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001*, 3-9. Ogunquit, Maine, USA: ACM. Accessed May 11, 2016. DOI: [10.1145/505168.505201](https://doi.org/10.1145/505168.505201).
- Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. "Data sharing by scientists: practices and perceptions." *PLoS ONE* no. 6 (6):e21101. Accessed May 11, 2016. DOI: [10.1371/journal.pone.0021101](https://doi.org/10.1371/journal.pone.0021101).
- Tenopir, Carol, Ben Birch, and Suzie Allard. 2012. *Academic Libraries and Research Data Services: current practices and plans for the future - An ACRL white paper*. Association of College and Research Libraries. Accessed May 11, 2016. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf.
- Wicherts, Jelte M., Marjan Bakker, and Dylan Molenaar. 2011. "Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results." *PLoS ONE* no. 6 (11):e26828. Accessed May 11, 2016. DOI: [10.1371/journal.pone.0026828](https://doi.org/10.1371/journal.pone.0026828).
- Wieczorek, John, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. 2012. "Darwin Core: an evolving community-developed biodiversity data standard." *PLoS ONE* no. 7 (1):e29715. Accessed May 11, 2016. DOI: [10.1371/journal.pone.0029715](https://doi.org/10.1371/journal.pone.0029715).
- Willis, Craig, Jane Greenberg, and Hollie White. 2012. "Analysis and synthesis of metadata goals for scientific data." *Journal of the American Society for Information Science and Technology* no. 63 (8):1505-1520. Accessed May 11, 2016. DOI: [10.1002/asi.22683](https://doi.org/10.1002/asi.22683).