Linked Data: An Oceanographic Perspective

Article *in* Journal of Ocean Technology · January 2013

CITATIONS
5

READS
172

5 authors, including:

Adam Marc Leadbetter
Marine Institute
70 PUBLICATIONS **369** CITATIONS
SEE PROFILE

Robert Arko
Columbia University
82 PUBLICATIONS **1,943** CITATIONS
SEE PROFILE

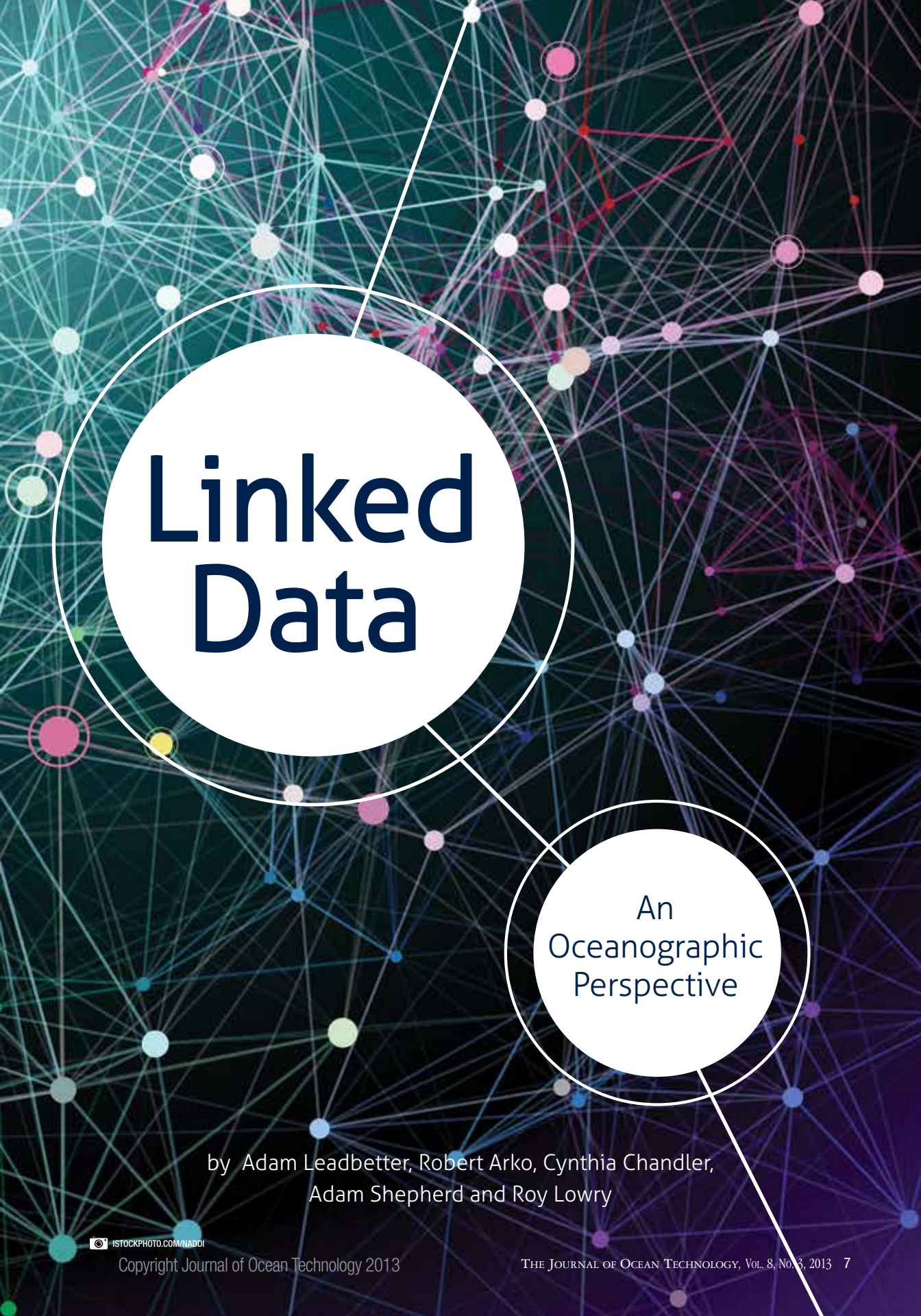
Cynthia Chandler
Woods Hole Oceanographic Institution
58 PUBLICATIONS **217** CITATIONS
SEE PROFILE

Adam Shepherd
Woods Hole Oceanographic Institution
19 PUBLICATIONS **90** CITATIONS
SEE PROFILE

Linked Data

An Oceanographic Perspective

by Adam Leadbetter, Robert Arko, Cynthia Chandler, Adam Shepherd and Roy Lowry



ISTOCKPHOTO.COM/NADDI

The fundamental design of the World Wide Web, familiar to anyone who has ever opened a browser, is a series of documents, images, and other media all accessible through the browser software. These documents are all identified by unique addresses (Uniform Resource Identifiers, or URIs) and those addresses are used to build links between the documents. In 2006, the principal architect of the World Wide Web, Sir Tim Berners-Lee, wrote a paper that uses these fundamental design tenets to formulate the idea of a Web of Data. The solution he proposed is commonly known as Linked Data[1]. Linked Data is built on four pillars:

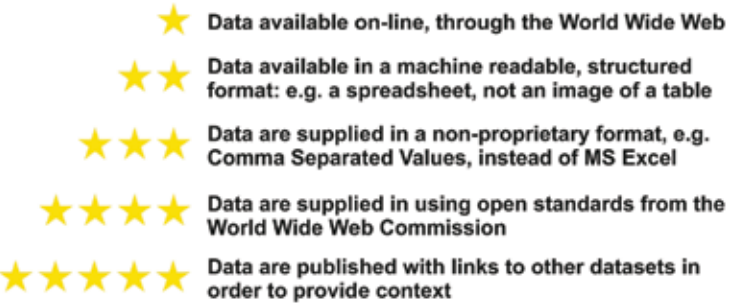
1. URIs are used to name data or metadata resources published on the World Wide Web

2. Users should be able to look up the URIs using a web browser (that is, the URIs use the HyperText Transfer Protocol, HTTP, so that they appear as web addresses)

3. The document to which the URI points should be encoded in a standard, as laid out by the World Wide Web Commission (W3C)

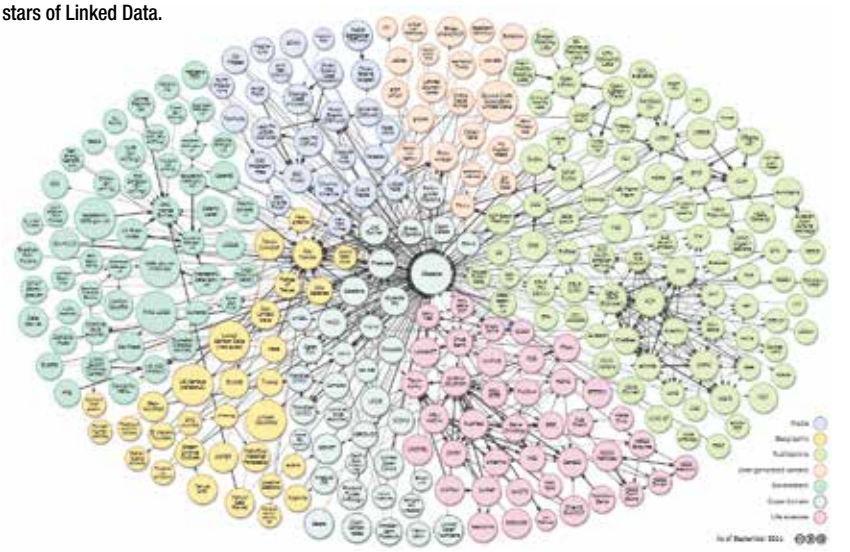
4. There should be links from one Linked Data document to the URIs of other Linked Data documents, so that users can discover more data or metadata (data about data) resources

In the seven years since the original Linked Data design document was released, a community of Linked Data publishers and users has grown up. The highest quality (Figure 1) published Linked Data has traditionally been represented by the Linking Open Data (LOD) cloud diagram (Figure 2). This diagram gives a high level overview of Linked Data datasets from around



SIR TIM BERNERS-LEE (1999)[1]

Figure 1: The five stars of Linked Data.



RICHARD CYGANIAK AND ANJA JENTZSCH. HTTP://LOD-CLOUD.NET/

Figure 2: Linking Open Data cloud diagram.

the globe, and it often features in presentations and documents on how data published on the World Wide Web can benefit society, especially if those data are linked together. However, at the time of writing, the Linked Open Data Cloud diagram appears to be no longer maintained. In a recent blog post, "The LOD cloud is dead, long live the trusted LOD cloud", Andreas Blaumer of the Semantic Web Company proposed the idea of domain specific, micro-Linked Data clouds. These micro-clouds should consist of resources used again and again by their specific domains due to the specific data or information presented and their highly active maintainers. In this essay, we present some of the recent developments that are leading to the formation of a Linked Ocean Data cloud.

Considering the 5-star rating for Linked Data as shown in Figure 1, it can be shown that many providers of oceanographic data are already publishing Linked Data to some level. At the lowest Linked Data levels, it is simply enough to make data files available online to data consumers. Improvements to the Linked Data star rating can be made through the use of non-proprietary data formats, such as Comma Separated Value files rather than, for example, Microsoft Excel specific spreadsheet files. Further improvements to the star rating may be made through more technical enhancements such as publishing the metadata or data using a W3C standard such as the Resource Description Framework. An example of three-star Linked Data in the oceanographic domain comes from the European Commission SeaDataNet project. This project publishes its Common Data Index metadata files online in an extensible markup language (XML) format, and these metadata files contain links to the associated data files which are presented in well-documented, non-proprietary formats.

To achieve the highest, 5 Star, Linked Data status, a dataset must be published using W3C standard formats, normally the Resource Description Framework (RDF). RDF is a data model that uses expressions of the form *subject-predicate-object*, known as triples. A simplistic triple would state *"the sea"* - *"has the colour"* - *"blue"*. A further 5 Star Linked Data requirement is that the resource must contain links to other Linked Data resources, so it must include triples of the form *"http://my/URI"* - "is related to" - "*http://your/URI*". A query language for RDF resources, known as SPARQL, can be used to interrogate resources. In the most recent specifications for SPARQL, it may be used to distribute queries across Linked Data resources allowing URIs from several hosts to be pulled together to form the resultset of a query.

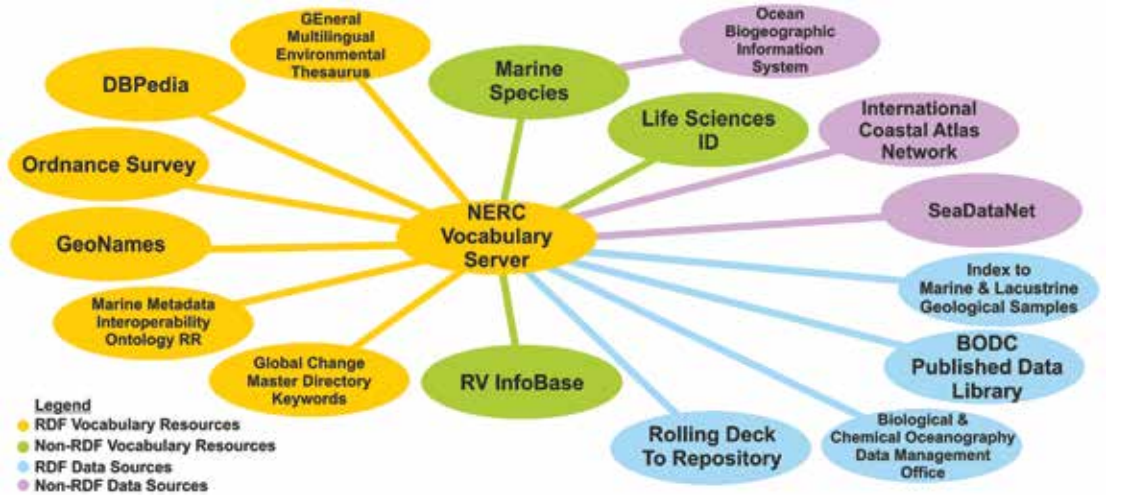
There are several existing 5 Star Linked Data resources published in the oceanographic domain. Two resources used to publish definitions of terms relevant to the marine domain are Version 2.0 of the Natural Environment Research Council (NERC) Vocabulary Server and the Marine Metadata Interoperability Ontology Registry and Repository (MMI-ORR). The NERC Vocabulary Server[ii] delivers content governed by a number of groups, including SeaDataNet, and a number of de facto standard vocabularies such as the RDF representation of the International Council for the Exploration of the Seas' Platform Codes. The MMI-ORR[iii] serves a number of vocabularies created by community and individual efforts, including those used within the United States Integrated Ocean Observing System and the Ocean Observatories Initiative. URIs from these resources are used to define parameters in data files in the Linked Ocean Data Cloud, and such metadata fields as: the sea area in which an observation was made; the data resources; the vessel or platform a measurement was made from; and the instrument used to make an oceanographic measurement. Many of the URIs on these two resources are interlinked, and are used within the 5 Star Linked Data datasets published by the United States National Science Foundation funded projects Rolling Deck to Repository (R2R) and Biological and Chemical Oceanography Data Management Office (BCO-DMO).

In the case of BCO-DMO[iv] and R2R[v], the two repositories manage related oceanographic data resources, such as research cruises and

their datasets, but from different perspectives with different goals. However, these two repositories, by linking their metadata to a common semantic resource such as the NERC Vocabulary Server, can now discover the other repository's related data through links to this common resource. Beyond the obvious advantage of increased resource discovery, this capability has made data management practices such as data validation, much easier to accomplish. For instance, using the SPARQL language, the two repositories can efficiently match their related cruise metadata to the other repository's cruise metadata through their common link to the NERC Vocabulary Server. Not only will a SPARQL query derive matches, but it can also be used to assess the accuracy of shared common metadata values. From these derived matches and assessments, data managers can quickly quality control their own cruise metadata based on the results from the other repository.

As described earlier, the Linked Open Data Cloud has been a popular visualization of the links between 5 Star Linked Data resources. In a similar manner, the Linked Ocean Data Cloud is an application of the ideas introduced recently by Andreas Blaumer, whose blog post presents an overview of the Clean Energy Linked Open Data Cloud and links to the German Library Linked Open Data Cloud. Executing a World Wide Web search for similar micro-Linked Open Data clouds quickly reveals the existence of a Linguistic Linked Open Data cloud. We have assembled the relevant Linked Data resources of which we are aware, and have developed the diagram that can be seen in Figure 3. Where possible, URI links have been built between resources in order to expand the scope of the diagram as much as possible at this early stage. The diagram indicates those resources currently publishing high level Linked Data in RDF and those currently publishing to a lower Linked Data level.

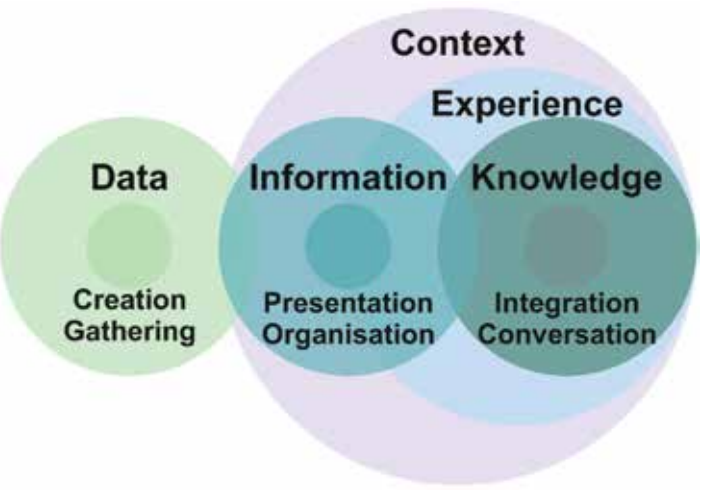
In order to assess the benefits of the Linked Data approach, we turn to a model taken from the domains of economics and education which has been used to promote and direct informatics research in recent years: the Data-Knowledge-Information "ecosystem" model (Figure 4). This model considers data to be the result of making observations; data are created or gathered in this way. By using metadata to provide context for the data, and presenting or organizing the data in a specific way information may be built up. In this case information is deemed to be something useful inferred from the observed data. Finally, using the contextual metadata and information, and



ADAM LEADBETTER

Figure 3: The Linked Ocean Data cloud.

the user's specific experience of their specialist subject area and more generally the world around them, knowledge may be gained. A 5 Star Linked Data resource provides extra contextual information for its components through the links to other Linked Data resources. Following these links also allows for new ways of presenting and integrating data and information, allowing new insights and new knowledge to be created using data which may previously have been unlinked. This means that adopting a 5 Star Linked Data approach to serving metadata and data is of benefit to researchers looking for new insights and for funding bodies who are looking to show additional impact of the data produced by their grants.



NATHAN SHEDROFF

Figure 4: The data, knowledge and information ecosystem.

All this is not to say that deploying a Linked Data resource is not without its difficulties. Deploying an effective Linked Data resource means that the URIs that are used to identify objects within the resource should be well designed in order to allow developers and clients to intuitively navigate the content. The United Kingdom government provides good advice on design patterns for URIs for Linked Data, which has been widely adopted in many countries. Another important consideration is the persistence of a URI, or the persistence of the data which the URI represents. If a URI becomes no longer valid, it is a fairly trivial matter for a software agent that is navigating the Linked Data cloud to ignore the response. However, if a human user has spent a long time building the links between one resource and another only to find that the URIs are no longer valid at one resource, the result is a degradation of trust. Or, if a person browsing through Linked Data resources discovers that several URIs in one namespace no longer represent anything, they are unlikely to trust that resource in the long term. In building these networks it is important to remember that while the approach may be rooted in information technology there are many issues derived from human values which must still be respected.

In conclusion, Linked Data is a decade-old general concept built on the foundations of the World Wide Web which has a simple set of best practices associated with it. The Linked Data paradigm is increasingly being seen as a driver in the next generation of collaborative research networks and scientific community building. This was demonstrated by Deborah McGuinness of Rensselaer Polytechnic Institute at the 2012 American Geophysical Union Fall Meeting in a presentation entitled "Community Science – The Next Frontier." While the overall concept of publishing Linked Data may be novel to many ocean science data providers, we feel that many are already offering Linked Data at some level and we hope they will aspire to join the Linked Ocean Data cloud and publish 5 Star Linked Data. Here we have shown that it is indeed possible to publish 5 Star Linked Ocean Data, and some of the benefits of this approach.

i http://www.w3.org/DesignIssues/LinkedData.html
ii http://vocab.nerc.ac.uk/
iii http://mmisw.org/orr/
iv http://linked.bco-dmo.org/
v http://linked.rvdata.us/

Dr. Adam Leadbetter is a data scientist at the British Oceanographic Data Centre, focusing on the development of controlled vocabularies for the marine science domain. In addition to his work in the realms of vocabularies and Linked Data, he has a particular interest in developing data into a citable entity for the purposes of scientific reproducibility and he is also co-chair of the technical committee of the International Coastal Atlas Network.

Robert Arko is a lead system analyst/ programmer at the Lamont-Doherty Earth Observatory, Columbia University. He is Technical Director of the National Science Foundation funded project Rolling Deck to Repository, which envisions the United States' academic research vessel fleet as an integrated global observing system, with routine underway data and documentation flowing directly from research vessels to a central shore-side repository.

Cynthia Chandler is an Information Systems Associate at Woods Hole Oceanographic Institution with over thirty years of experience of integrating information systems technology with oceanography, including 35 research cruises. All of her current research is related to improving and enhancing data management activities throughout the full research data lifecycle, from 'proposal through preservation.' A particular focus is using semantics to connect resources to improve discovery, facilitate interoperability and ultimately make ocean science research data, especially data acquired during research cruises, more accessible to the research community.

Adam Shepherd is an Information Systems Associate at Woods Hole Oceanographic Institution with thirteen years of experience of building web-enabled software for the ocean sciences. Currently, his research focus has been on Linked Data applications, including vocabulary mapping tools, with a concentration in open source technologies.

Roy Lowry is the technical director of the British Oceanographic Data Centre with over thirty years experience of marine data management. He was recently awarded the International Oceanographic Data and Information Exchange Achievement Award for his contributions to file format standardization and semantic technology developments in the field of marine science.