New Ideas for Communities of Practice: Networks of Networks

Wim Hugo

South African Environmental Observation Network, Persequor Park, Pretoria, 0001.
wim@saeon.ac.za

Abstract

The last decade has seen the emergence of two interlinked trends in the support that information technology can bring to Earth and Environmental Observation Systems. These are, broadly speaking, focused on the increasing use of discoverable, brokered data and services, based on interoperability specifications, and the emergence of collaborative portals to support the production and use of research and development output.

There are three major problems with these developments, despite its benefits: (1) ***scalability*** *– the infrastructure is resource-intensive in respect of maintenance and extension (2)* ***efficiency*** *– a large part of the knowledge embedded into individual projects, initiatives, and collaborations are lost or not adequately captured by traditional meta-data, and (3)* ***flexibility*** *– current approaches are not designed for a movement to massive use of social networking, mobile and other devices that are connected to the internet, and similar developments.*

In this paper, we present new concepts that support a 'Knowledge Network of Networks' that can adapt to changing technologies, is self-maintaining and scalable, and can be supported by a variety of clients in any number of interaction channels – from traditional desktops and laptops to mobile phones, and smart devices. We have used the existing South African Earth Observation System of Systems (SAEOSS), South African Environmental Observation Network (SAEON), and South African Risk and Vulnerability Atlas (SARVA) portals as a case study for exploring some of the implications of these new concepts, especially for self-maintaining communities of practice. The paper then sets out a vision for a Reference Model for Scalable Knowledge Networks.

The underlying fabric of the concept is a massive, open, and liberalised meta-data resource that can be mined for new information and knowledge, and serves as a record of scientific endeavour.

1. Introduction

The concept of knowledge networks as a formal construct expressed as a collection of relationships between entities is not new, and has been applied to construct graphs of scientific endeavour in the recent past (Boyack, Klavans, and Börner, 2005), (Börner, 2011). These efforts were based on relationships inferred from journal publications and citations, and, as such, did not address what can be referred to as 'The Dark Matter of Science' (Uhlir et. al. 2013), largely referring to the relationships and knowledge available through data centres and services (implying meta-data), and the tacit knowledge embedded into the non-formalised record of scientific activity: project and institutional websites, funder databases, community of practice portals, and social networks. An addition, past efforts did not have access to the growing availability of persistently identifiable data objects in the web, as driven by the concept of 'Linked Open Data' (Linked Data, 2013).

Communities of practice often aggregate and disseminate their collective knowledge as a portal or science gateway. The extent to which they can be supported by knowledge networks is not a novel idea (Allee, 2000), but irrespective of the sources of data to create such a knowledge network, and its scope of application, it usually suffers from three interrelated problems: (1) **scalability** – the data infrastructure to support knowledge networks is resource-intensive in respect of maintenance and extension (2) **efficiency** – a large part of the knowledge embedded into individual projects, initiatives, and collaborations is not adequately captured by traditional meta-data, and (3) **flexibility** – current approaches are not designed for a movement to massive use of social networking, mobile and other devices that are connected to the internet, and similar developments, since meta-data standards are rigidly defined and while extensible, do not make provision for capturing contributions from multiple sources.

Within data and service portals set up by communities of practice, the general problems of scalability, efficiency, and flexibility in 'systems of systems' based on aggregated meta-data can be further defined in three use cases:

- **Use Case 1 – Coverage and Scope:** A way must be found to accurately assess the scope and coverage of and rationale for new projects, grants, data sets, and network initiatives. The alternative is lack of coordination, duplication, and waste of scarce resources (Efficiency).
- **Use Case 2 – Complexity and Efficiency**: An optimum level of meta-data detail can be defined, that is less formal and more elaborate than current standards-based meta-data, yet more efficient than semantic web solutions. To make Knowledge Networks scalable and self-maintaining, one of the requirements will be to enlist the growing possibility of small contributions by many, rather than large contributions by the few (Efficiency, Scalability, Flexibility).
- **Use Case 3 – Scalable, Agile Networks:** Given that formal meta-data can be supplemented by simple, standards-aligned additions from a wide variety of sources, it will be possible to create, maintain, and extend the supporting fabric of knowledge networks in a scalable and agile manner (Scalability, Flexibility).
- **Standards and Specifications**: To guide the widespread implementation of the knowledge networks needed to support the concept, the ideal pathway would be the development of a Reference Model for a 'Knowledge Network of Networks' that supports these use cases. It is highly unlikely that we need new sets of standards and specifications on a wide scale to implement such a reference model - most of the work will be selection and endorsement of existing resources, supplemented by reference implementations and compliant tools.

2. Mandates and Rationale

The world of science is saturated with projects, collaborations, special networks, etc. – collectively we call these "***Initiatives***" that can be characterised in a number of ways, including the extent to which important ***dimensions*** are covered by each initiative. These dimensions are, for example, aspects such as coverage (temporal, spatial, and topic/ semantic coverages), scalability of meta-data, degree of structure of the information objects, and so on. Using this approach, it is possible to ***characterise*** and ***compare*** initiatives (A study is required to identify a minimum set of orthogonal dimensions, and the examples do not claim to be these). The two figures below show, respectively, examples of how such dimensions can be defined, and compares two initiatives (The Group on Earth Observations (GEO, 2013), and The ICSU World Data System (ICSU-WDS, 2103)) using the examples.

Building a repository of the "initiatives" in science, including network initiatives, in this way enables us to do two things in broad terms:

1. Assess the overlap amongst and gaps between initiatives;
2. Determine to what extent current dimensions should be broadened to allow maximum utility to users.

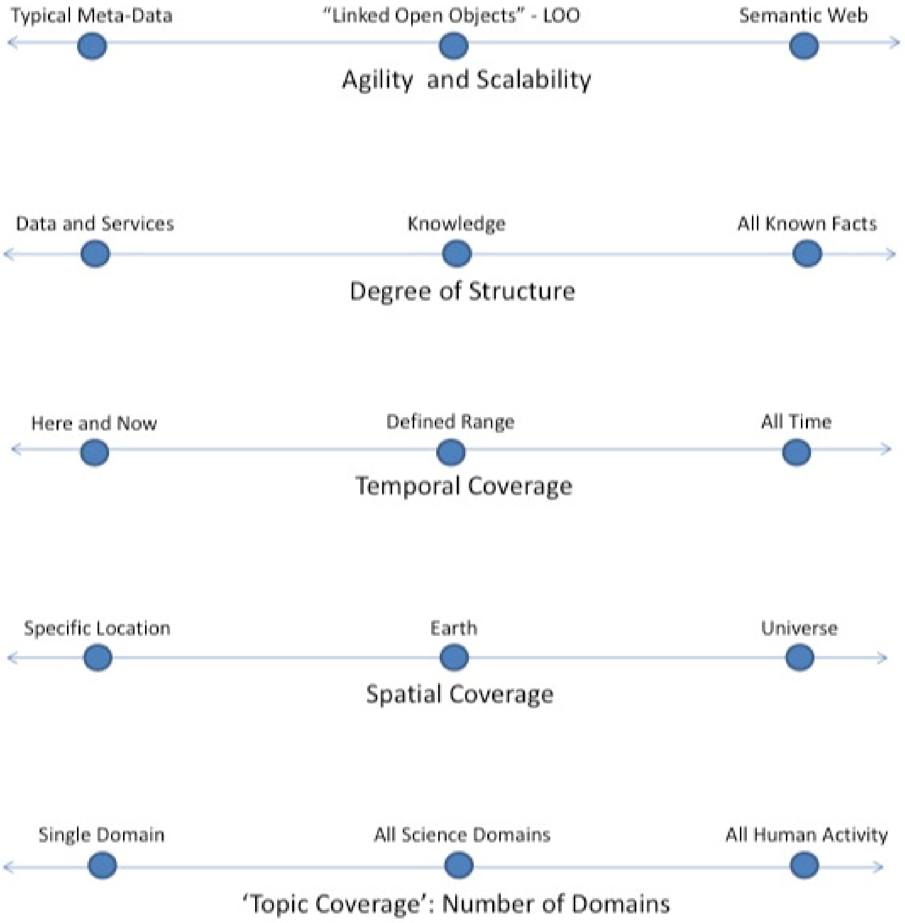

Figure 1: Example of a 5-Dimensional Characterization, with simple scale elements

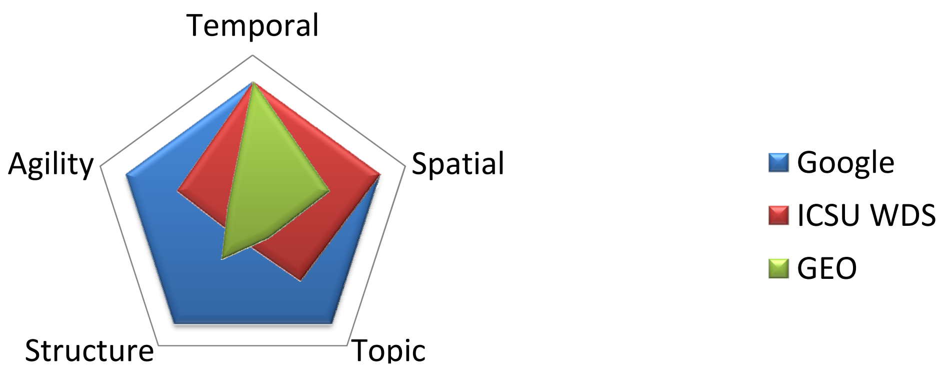

Figure 2: Comparing Initiatives using Dimensions

Practical realisation of this concept will require the following:

- Evaluate and define the mandate dimensions that distinguish networks and network initiatives from one another.
- Assess the mandates and coverage of network organisations in terms of these dimensions.
- Provide tools and services for new initiatives and funding agencies to evaluate gaps, overlaps and duplication.
- Study the extent to which mandates must be broadened and structures liberalised to maximise utility.

This maximisation of utility can be viewed graphically in Figure 3:

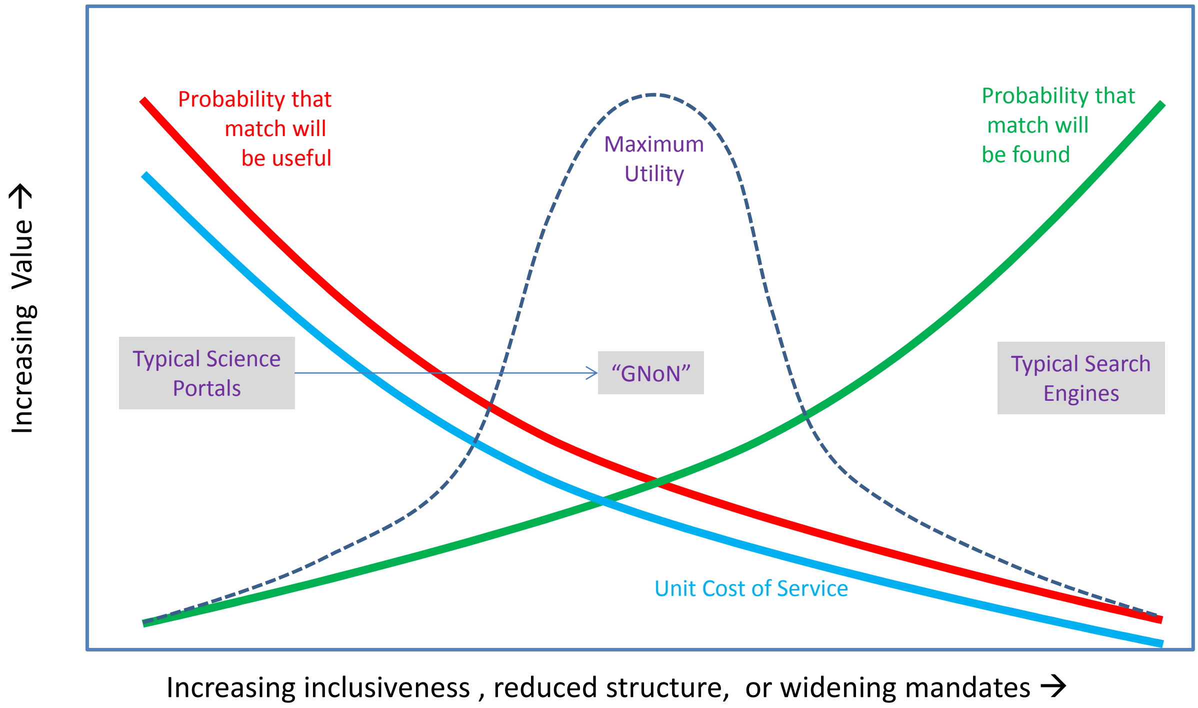

Figure 3: Maximising Utility by Designing Scope

In our opinion, typical science portals usually have a narrowly defined scope, focused on traditional meta-data and structured data for a specific discipline - which does not maximise utility for the user. On the other hand, a typical search engine and the availability of a fully semantic web (W3C, 2013) operate at the opposite end of the scale: it references a large number of resources with little context.

Part of the challenge for initiatives to establish the infrastructure will be to find, possibly through trial and error, to what extent typical science portals must be extended and liberalised to improve utility. This discussion ties into the next use case: that it is possible to manage complexity and find an optimal, practical way to work with linked objects in the web.

3. Complexity and Efficiency

The worldwide web, if extended to define all of the information or knowledge objects that can be found at each physical address, will be large and complex. If we design any arbitrary number of additional semantic relationships between these nodes, the complexity grows more or less as the square of the number of nodes (i.e. asymptotically) *for each additional relationship*. This is not a feasible long-term situation in terms of information retrieval and management (Fensel and van Harmelen, 2007,).

Hence we are confronted with two extremes:

- **The Complete Web**: every piece of information at a physical network node is potentially in multiple relationships with every other. This enormous network (the "Semantic Web (W3C, 2013)") is many times larger than the physical internet and is ***unlikely to be practically useful***.
- **Formal Meta-Data**: very limited links are formally specified, eliminating almost all of the potential links between pieces of information to favour only a very rigid collection.

Figure 4 provides a sense of the enormity of the problem:

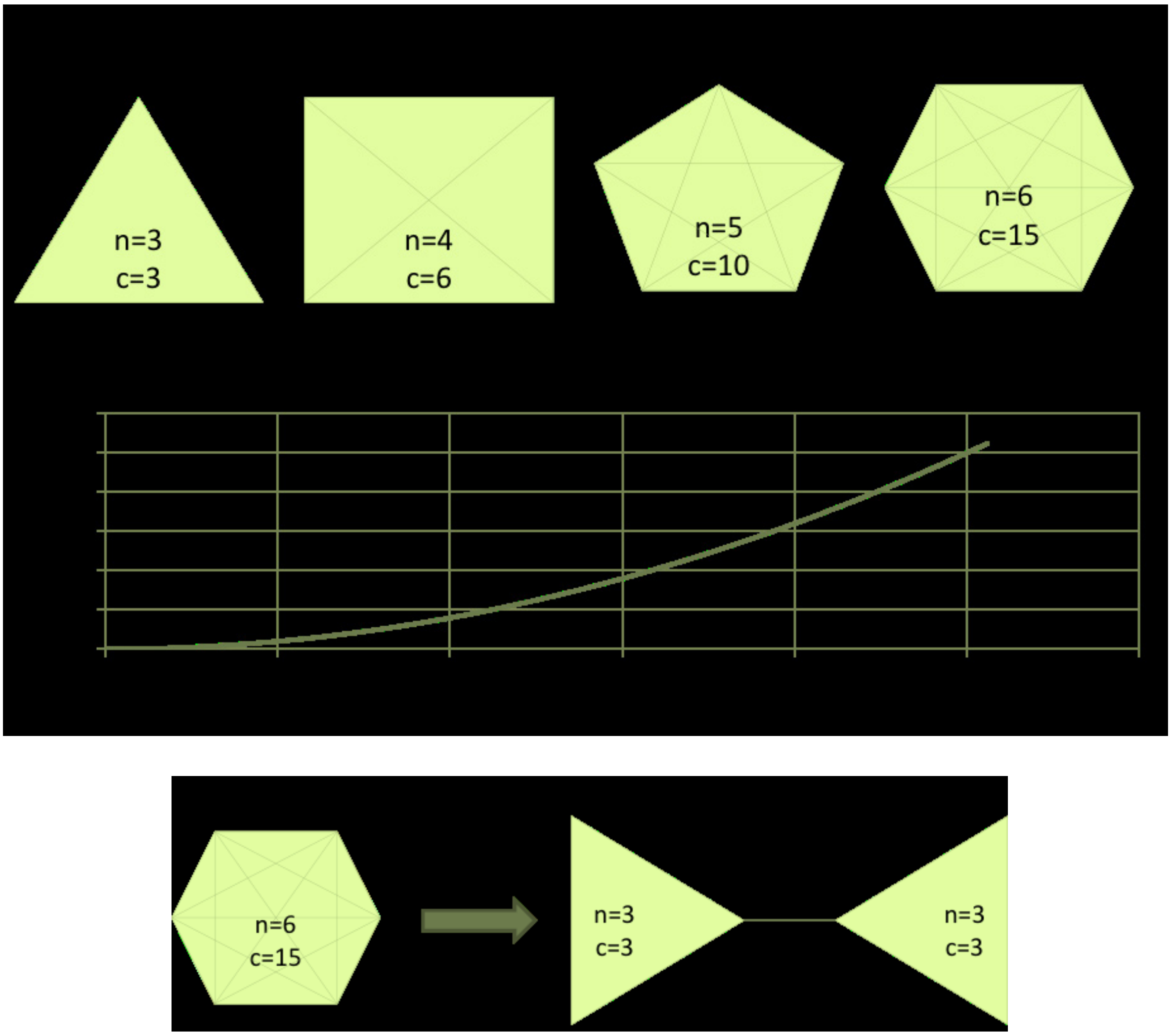

Figure 4: Complexity and Relationships

As the figure partly illustrates, there will be generic strategies for reducing the complexity:

1. Clustering: measures to classify into levels of detail and dependency - achieved through controlled vocabularies, ontologies, and thesauri, and elimination of non-essential, redundant, or duplicate relationships;
2. Approximation or Sampling Scale: Some relationships are stronger than others – we can disregard the weakest to obtain an approximation or less accurate network.

Practical implementation will require ***networks that are just complex enough*** to serve the purpose:

- Define a conceptual model for the aspects of knowledge networks that we want to manage.
- This fixes the relationships between types (RDA, 2013) in such networks: people, institutions, funders, projects, collaborations, data, services, publications, etc. By eliminating non-essential relationships, complexity is reduced.
- The conceptual model extends meta-data standards, and one should actively promote the formal and informal contributions to such an extended, distributed resource.

- It will be highly beneficial if multiple global initiatives can maintain repositories or (registries) of network-defining extended meta-data as a public good.
- The conceptual model can draw on and extend resources such as schema.org (Schema.org, 2013).
- The practical implementation will continuously benefit from the development of appropriate ontologies, thesauri, and controlled vocabularies.

4. Scalable, Agile Networks

The typical current network is persisted in a portal-type application; and in such applications the majority of knowledge is static: the relationships in the portal defining participating institutions, people, the aligned projects and tasks, and the funding partners are typically indirectly available in such a portal and 'hard-coded' – in effect, a 'Community of Practice'. Figure 5 shows an example of such a portal, based on assessment of the SAEON Data Portal, the South African Earth Observation System of Systems, The South African Risk and Vulnerability Atlas, and the prototype World Data Centre for Biodiversity and Human Health in Africa.

In some cases, it is likely that dynamic resources (such as lists of publications and data applicable to the network or initiative, tools and software, or other knowledge objects) are available, but typically linked into the portal environment manually.

It is increasingly common for some resources, such as data links or meta-data listings, to be automated, but in most cases these resources are not always dynamically updated and still require some form of manual intervention to maintain. The same is true of services. (We exclude a class of content that can be seen as true context, such as user-contributed articles or discussions possibly linked to automated resources).

There are at least two problems with these traditional portals:

- The information and knowledge encapsulated in the portal is not easily re-usable in another context, and critically, in many cases ***do not survive the termination of the contract or project*** that led to the creation of the network.
- It is time-consuming, and expensive, to create and maintain these portals. The manpower typically employed for such environments are domain ***and*** information technology experts or 'data scientists' – difficult to find, keep, and remunerate (TIBCO, 2012).

In contrast, it is possible to envisage a future situation where we will need a mixture of only two tool-driven processes: human mediation of dynamically updated resources, and automated mediation of such resources. This possible arrangement is shown in Figure 6.

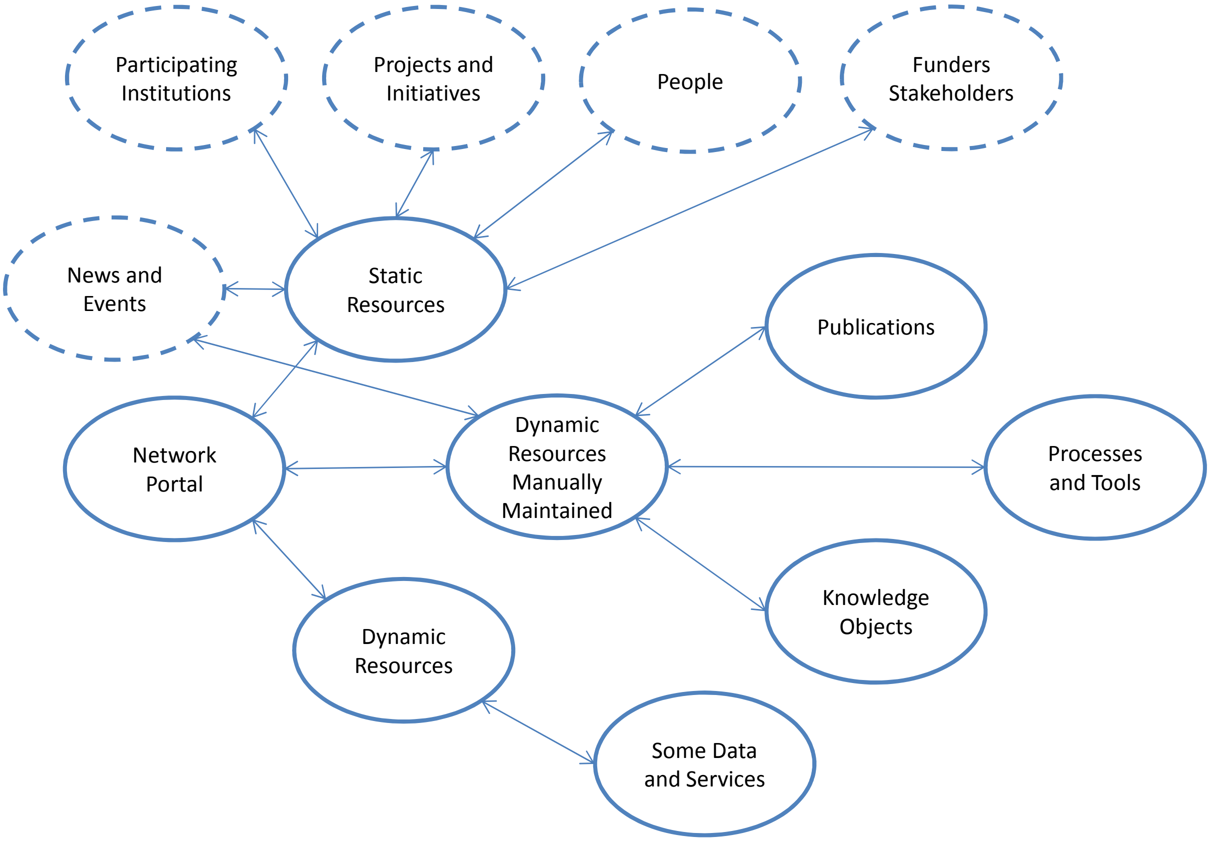

Figure 5: Typical Current Network Encoding as a Portal Serving a Community of Practice

By ‘mediation’, in this context, we mean the process of evaluation and ranking or filtering of raw resources available from the ‘networks of networks’ type repositories described earlier. The typical content or knowledge that can be mediated or judged by automated means include data, services, processes, tools, and (structured) knowledge objects. Tools for automated mediation of these resources are typically easy to construct, based on metrics such as rankings, costs, speed of access, or license conditions.

Human judgment will, of course, be required, but if the resources in the repositories are available in a standardised way (e.g. Resource Description Framework triplets (RDF, 2013)), then it is again a relatively simple task to construct assistant tools whereby value judgments can be applied to dynamically changing resources as a set of rules. Furthermore, in a distributed environment where services and information objects exhibit signatures that are predictable, ***tests become a practical reality*** as a means of automated evaluation.

In this way, it will be possible to construct multiple (competing or complementary) client applications based on the standardised repository that allow different viewpoints into the available knowledge network: institutional viewpoints, personal viewpoints, project viewpoints, topic viewpoints, community viewpoints, and the like. It is already possible, using the increasingly rich availability of visualisation services, to create standard views of organisation structures, people networks, density maps, tag clouds, and so on, based on repository queries as feeds.

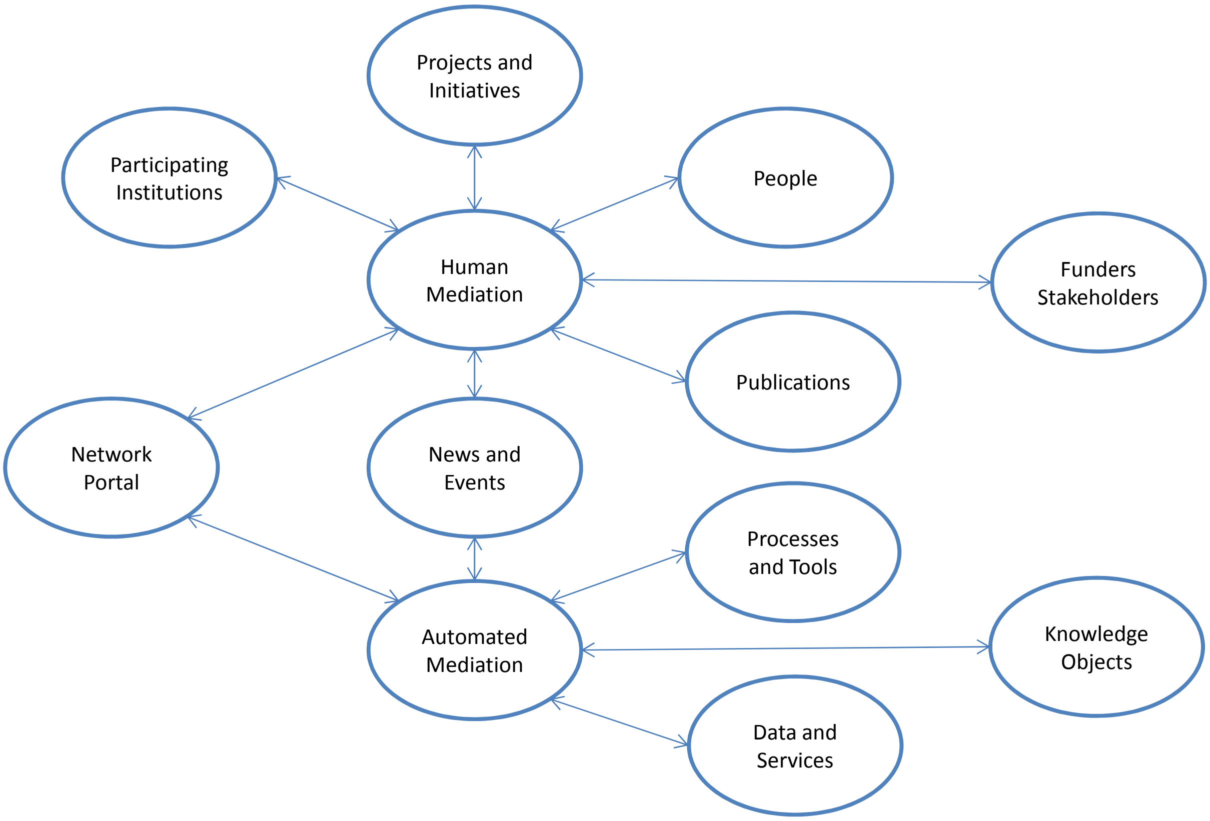

Figure 6: Self-Organising, Scalable Future Portal to Serve a Community of Practice

Figure 7 describes a detailed use case with 5 variations. In this use case, we are

- Mining meta-data to build an inventory of links between collaborators,
- Updating a science blog to note that a person has a new collaborator,
- Pushing data from the meta-data mining activity to an RDF repository,
- Harvesting relationships from a science blog conforming to expected standards into the RDF repository,
- Querying the RDF repository with a general purpose tool to look for potential collaborators,
- Or query collaborators of a given person automatically and maintain the links in a web page in a portal dynamically.

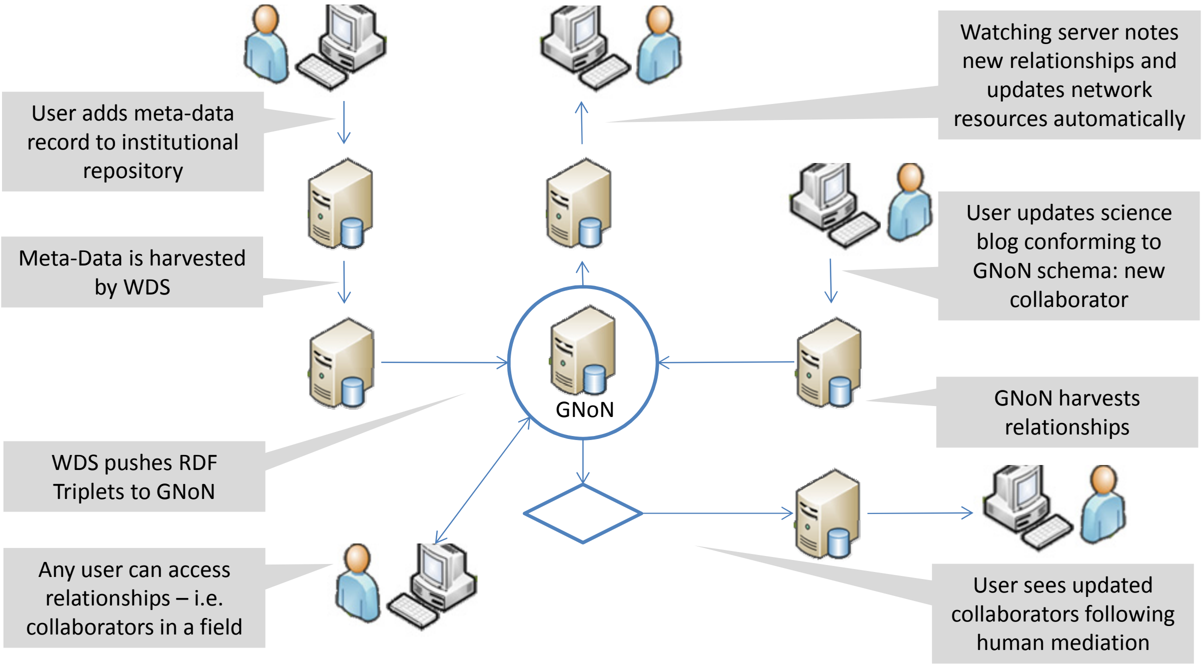
User adds meta-data record to institutional repository
Meta-Data is harvested by WDS
WDS pushes RDF Triplets to GNoN
Any user can access relationships – i.e. collaborators in a field
GNoN
Watching server notes new relationships and updates network resources automatically
User updates science blog conforming to GNoN schema: new collaborator
GNoN harvests relationships
User sees updated collaborators following human mediation

Figure 7: Simple Use Case

As illustrated in Figure 8, some 'Global Network of Networks' (GNoN) can assist in this process as follows:

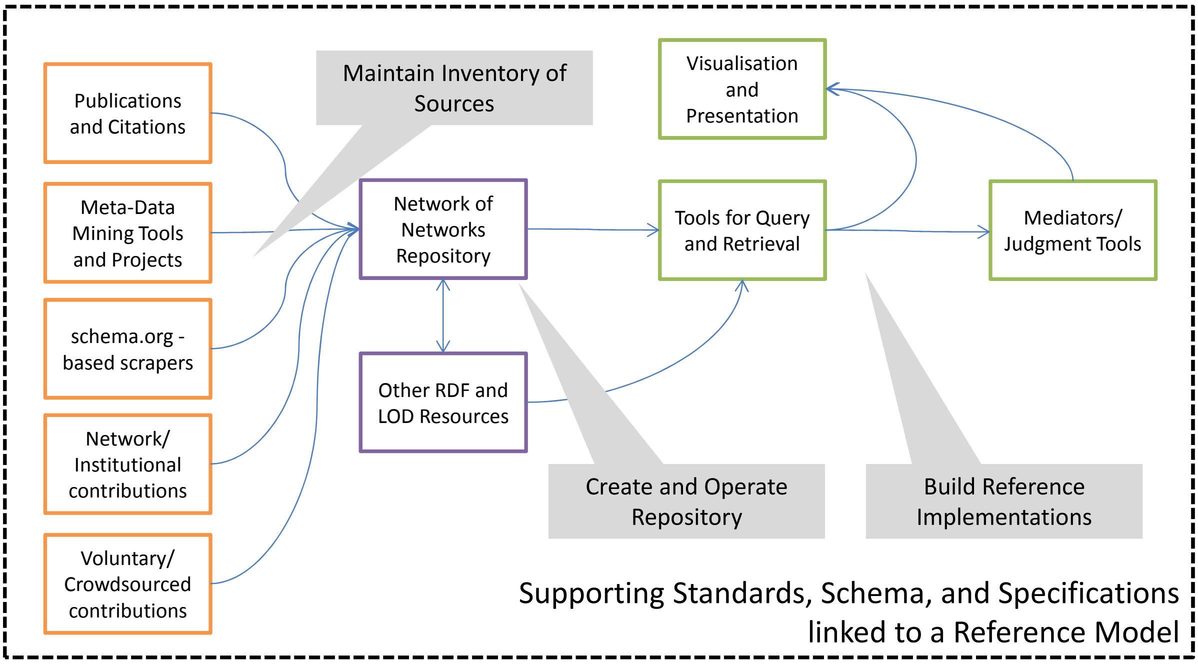
Publications and Citations
Meta-Data Mining Tools and Projects
schema.org - based scrapers
Network/ Institutional contributions
Voluntary/ Crowdsourced contributions
Maintain Inventory of Sources
Network of Networks Repository
Other RDF and LOD Resources
Visualisation and Presentation
Tools for Query and Retrieval
Mediators/ Judgment Tools
Create and Operate Repository
Build Reference Implementations
Supporting Standards, Schema, and Specifications linked to a Reference Model

Figure 8: Building a new Meta-Data Resource

1. By maintaining a 'Networks of Networks' repository, using at least the following methods:
    a. Accessing database services on publications and citations,
    b. Mining existing meta-data repositories,
    c. Scraping existing sites for meta-data based on extensions to the schema.org concept (Schema.org 2013),

   d. Accepting contributions from networks, institutions, initiatives, and the like,

   e. Allowing voluntary contributions from blogs, social media, Virtual Research Environments (VREs), and similar sources.

2. Enhancing this repository with appropriate relationships to existing and emerging Linked Open Data (LOD) (Linked Data 2013) repositories.
3. Building reference implementations of Query, Mediation, and Visualisation tools.

5. Conclusion: Reference Model, Standards, and Specifications

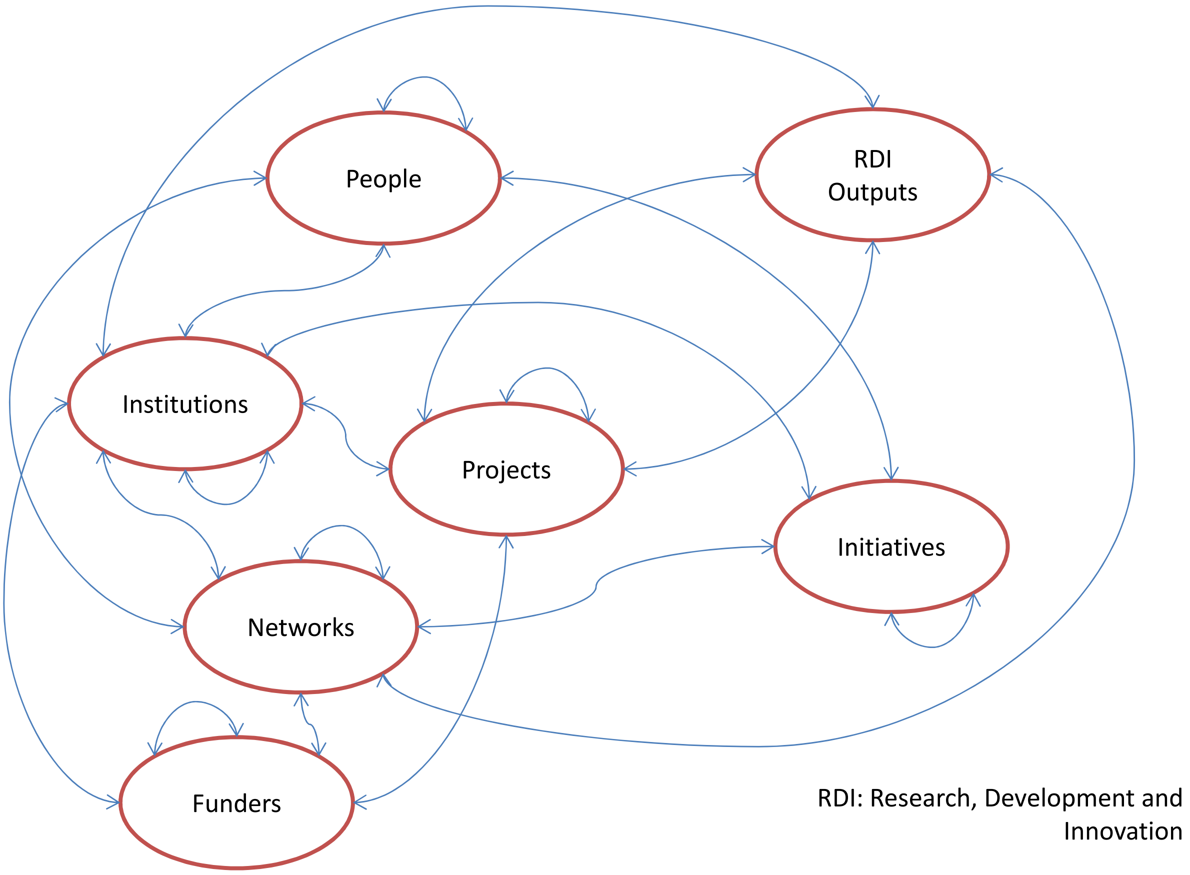

Figure 9: Example of a Conceptual Model for the Main Entities implied by Funded Research

We believe that it will be possible to define and support a ***Reference Model for Scalable Knowledge Networks*** without having to create a large number of standards and specifications. In broad terms, the following families of standards will be required:

1. A ***conceptual model*** needs to be defined (see above), this allows extensions to formal schema so that it is possible to create web resources that are harvestable as ‘relationship’ meta-data.
2. The schema extensions, should they not exist already in publicly maintained resources and type registries such as schema.org, could be added to these to make it universally accessible.

3. It is likely that the primary encoding for the harvested relationships will be based on the Resource Description Framework (RDF), making it functionally similar to the Linked Open Data initiative.

If we consider traditional meta-data, of which an example is shown in Figure 10, it is clear that the structure is mostly hierarchical, and that obtaining relationships between nodes in the hierarchical 'tree' will be difficult, or will not be supported by the information contained in these meta-data records.

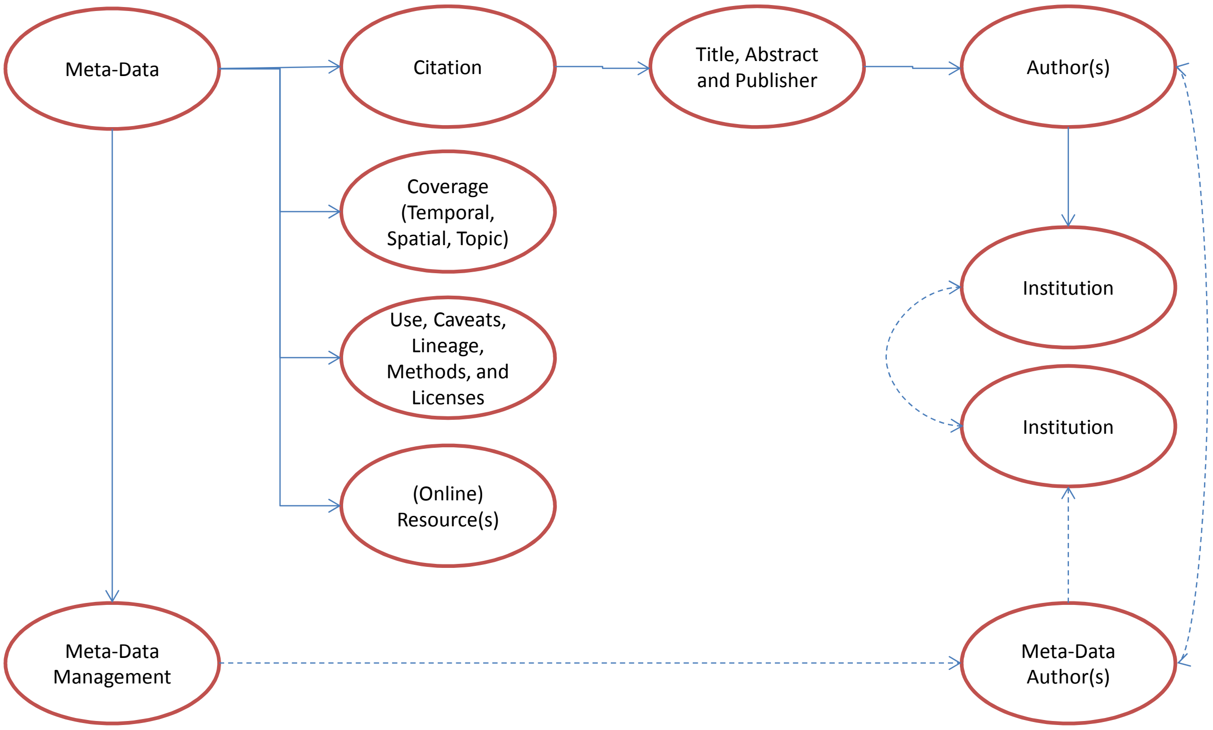
Meta-Data
Citation
Title, Abstract and Publisher
Author(s)
Coverage (Temporal, Spatial, Topic)
Institution
Use, Caveats, Lineage, Methods, and Licenses
Institution
(Online) Resource(s)
Meta-Data Management
Meta-Data Author(s)

Figure 10: Typical Traditional Meta-Data Record (Based on Assessment of ISO 19115, SANS 1878, EML, and Dublin Core)

In contrast to this arrangement one envisages a more liberalised meta-data schema, in which contributions to a distributed repository of such meta-data can be obtained in a variety of ways:

1. ***Mining existing meta-data records***: Existing aggregations of formal meta-data provide a rich starting point for mining the basic relationships embedded into the meta-data implicitly.
2. ***Websites conforming to Schema***: if websites for the institutions, networks, projects, and people involved in scientific output production conform to known schema, such as schema.org, they can be mined for additional meta-data, reinforcing and extending the relationships that are available. (An excellent example is provided by rNews.org (IPTC 2013).
3. ***Virtual Research Environments and Social Media*** are likely to become a more integral part of scientific endeavour in future: recording collaborations, methods, blog-style assessments of resources, etc. These can be pushed to meta-data repositories or mined for additional relationships.

4. ***Membership Data*** in existing network organisations form a specific and valuable base of knowledge about the state of networks at any given time.
5. ***Ontologies and vocabularies*** are used to cluster the relationships, and to reduce the complexity of the resulting network of meta-data. Other methods may include metrics such as weight of use, rankings, costs, etc.
6. ***Unification of Meta-Data Standards:*** A desirable side-effect of this approach is that it serves as a centralised cross-walk between divergent meta-data standards.

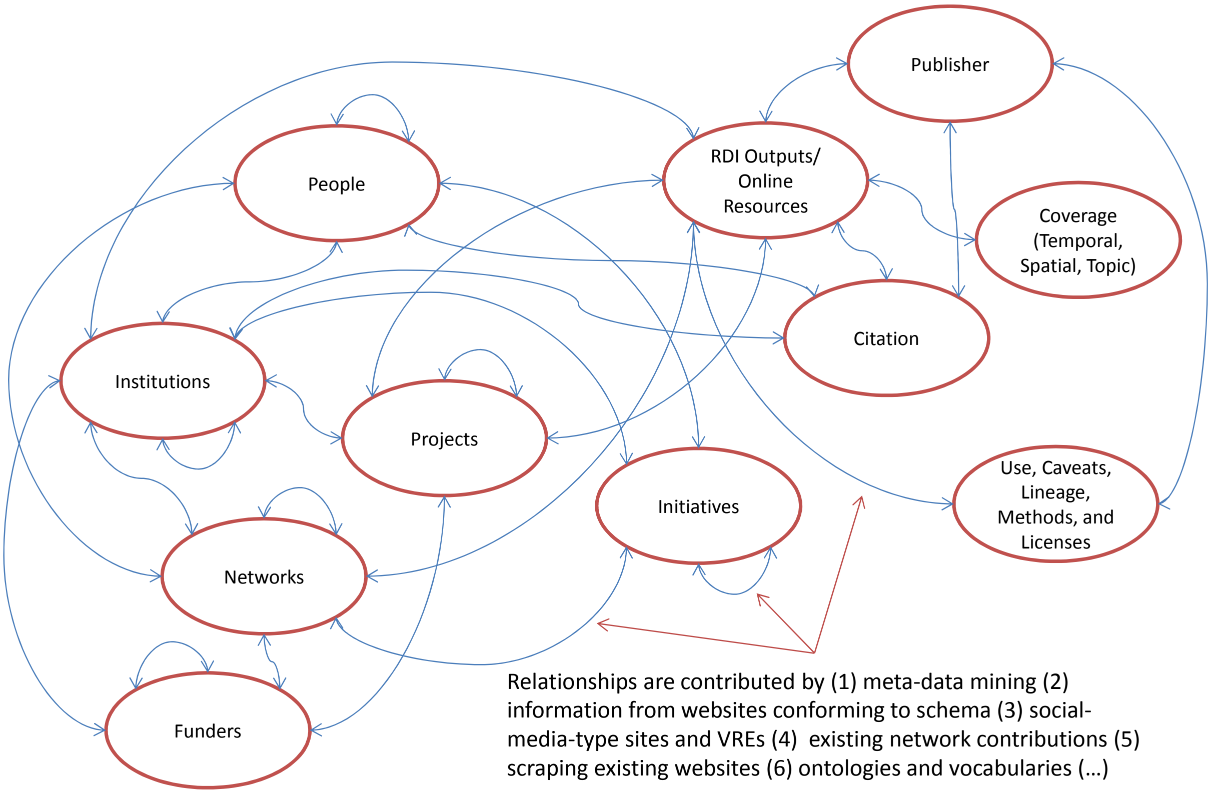
Publisher
People
RDI Outputs/ Online Resources
Coverage (Temporal, Spatial, Topic)
Citation
Institutions
Projects
Initiatives
Use, Caveats, Lineage, Methods, and Licenses
Networks
Funders
Relationships are contributed by (1) meta-data mining (2) information from websites conforming to schema (3) social-media-type sites and VREs (4) existing network contributions (5) scraping existing websites (6) ontologies and vocabularies (…)

Figure 11: 'Liberalised' meta-data with contributions from a variety of sources

For the development of such a reference model, the likely tasks include:

1. Development of an architecture supported by a conceptual model;
2. Identification of important patterns in the architecture;
3. Identifying, or developing, standards, specifications, profiles, and protocols to support the model;
4. Developing compliant reference implementations and tools;
5. Maintaining the governance framework for the initiative;
6. Building capacity and champion the adoption of the initiative.

6. Acknowledgement

The work reported in this paper is partly supported by NRF through the author's involvement with the ICSU World Data System.

7. References

Allee, V. (2000). Knowledge Networks and Communities of Practice, OD Practitioner, Fall/Winter 2000. Viewed April 2013, http://www.vernaallee.com/images/VAA-KnowledgeNetworksAndCommunitiesOfPractice.pdf

Börner, K. (2010). *An Atlas of Science*. Cambridge: MIT Press.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*(3), 351-374.

Dublin Core (2013). International Standard for General-Purpose Meta-Data, Viewed January 2013, < http://dublincore.org/>

Fensel, D. and van Harmelen, F. (2007). Unifying Reasoning and Search to Web Scale, IEEE Computer Society, 1089-7801/07. http://www.cs.vu.nl/~frankh/postscript/IEEE-IC07.pdf

GEO (2013). Group on Earth Observations, Viewed January 2013, http://earthobservations.org/index.shtml

ICSU-WDS (2013). International Council of Scientific Unions – World Data System, Viewed January 2103, http://icsu-wds.org

IPTC (2013). Introduction to RDFa, Viewed January 2013, http://dev.iptc.org/Introduction-To-RDFa

ISO 19115 (2013). International Standard for Geospatial Meta-Data, Viewed January 2013, http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020

KNB (2013). Environmental Metadata Language, Viewed January 2013, http://knb.ecoinformatics.org/software/eml/

Linked Data (2013). Connect Distributed Data Across the Web, Viewed January 2013, http://www.linkeddata.org

RDA, 2013. Data Type Registries Working Group, Viewed April 2013, http://forum.rd-alliance.org/viewtopic.php?f=2&t=30

Schema.org (2013). (A Collaboration Between Major Search Engines), Viewed January 2013, http://www.schema.org

TIBCO (2012). How to Battle the Looming Shortage of Data Scientists, SpotFire Blog, Viewed April 2013, http://spotfire.tibco.com/blog/?p=10429

Uhlir, P et. al. (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology – CODATA-ICSTI Task Group on Data Citation Standards and Practices, CoDATA, Paris (In Publication)

W3C, 2013. W3C Semantic Web Activity, viewed January 2013, http://www.w3.org/2001/sw/