01/07/2013

WDS Knowledge Network Concept: Summary[1]

Contributions from: Kim Finney, Michael Diepenbroek, Rob Atkinson, Peter Fox, Jane Hunter, Mustapha Mokrane, Yasuhiro Murayama, Mark Parsons, and Wim Hugo.

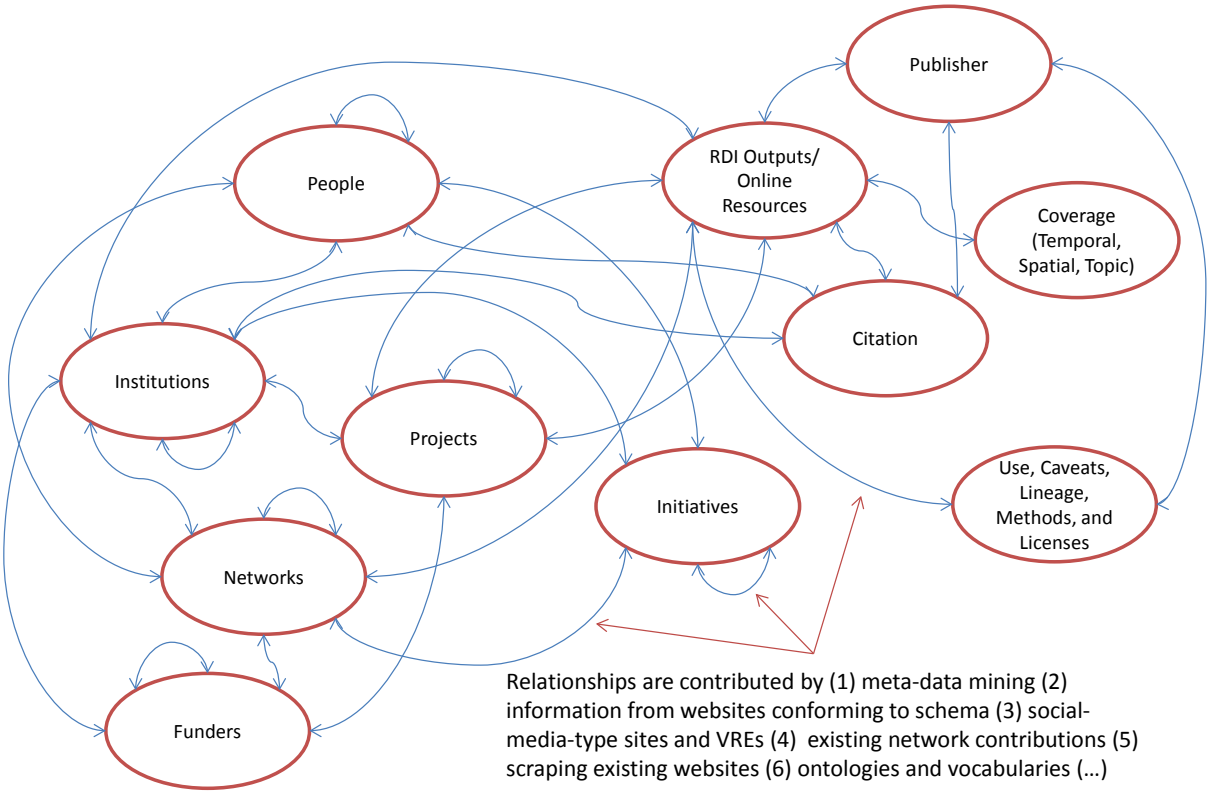
The ICSU World Data System (ICSU-WDS) has a stated objective to establish a metadata aggregation from the metadata holdings of its membership so that the extent of data and service offerings across this membership can be discovered through a single interface. This task will entail technical and semantic integration challenges because of the different terminologies, vocabularies, and information schemas being used. Nevertheless, once it is possible to view the content and services available from WDS Members, significant value will be added for Members and stakeholders alike.

ICSU-WDS would like to go further than merely providing a single window into its Members' metadata, however; it would also like to create a broader Knowledge Network. In summary, the vision for this Network is that additional information sources will be used to supplement Members' metadata and will be linked to them through the use of Linked Data.

Linked Data simply means identifying specific concepts (e.g., a project, a researcher, an address, a funder) that can be denoted by a unique resource identifier, which can then be referenced and inspected through a Hypertext Transfer Protocol request. Through this extension into Knowledge Networks—constructed mainly from the aggregated metadata repository and the WDS International Programme Office's (WDS-IPO's) existing database of Member information, and by integrating these with existing Linked Open Data repositories—ICSU-WDS feels that it can significantly enhance the usefulness of its metadata resource and improve the ability of researchers, funding agencies, and data centre management to understand and apply this information.

The Knowledge Network is a Web-based, interlinked repository of relationships between the actors and entities that make up our research landscape: people, institutions, projects, research disciplines and topics, funding sources, and the like. Linked Data are usually expressed in the Resource Description Framework (RDF) language, and the relationships between linked concepts are constructed through the use of triples (i.e., statements containing a subject, a predicate, and an object; for example, the statement 'the researcher has a publication' consists of 'the researcher' as the subject, the relation 'has a', and an object of 'publication'). The types of entities that we would be seeking to relate, describe, and expose to users are summarized in the below diagram.

[1] Discussion based in large part on Hugo, W. (2013): "New Ideas for Communities of Practice—Networks of Networks", SAJG, Vol. 2, No. 2, 2013 – http://www.sajg.org.za/index.php/sajg/article/view/96



In time, the metadata and WDS Member database resources can be supplemented by other sources of information drawn from social networks, volunteered or crowd-sourced information, and so on; but our initial focus is on formal data obtained from metadata and organizational details provided by participating WDS Members.

During the 2013 initial implementation process, ICSU-WDS will thus focus on only metadata repositories contributed by its Members and available membership data. However, other potential information sources will also become operational in due course and will contribute to a more comprehensive view of the scientific landscape.

The scope of work and the components that are needed to create an operational prototype of the WDS Knowledge Network are broadly expressed as the following:

1. We need to understand the nature of WDS Members' metadata and determine the common entities and concepts that are provided routinely to ICSU-WDS as part of the membership application process. This constitutes the material that is readily accessible for utilization.
2. On the basis of that information, we need to then create an extensible conceptual model capturing the actors and entities—as well as the relationships between them—that we would like to use to seed the Knowledge Network. This step also defines the scope of the triplets (RDF statements) that we will be required to ingest, and a suitable RDF-based triple store must be established to manage the statements.

3. We need to also identify credible and persistent Linked Open Data concept repositories and content sources for the entities and actors (people, institutions, citations, etc.) that will be reference in the Network.
4. A process of data mining must be eventually applied to WDS Members' metadata collections and the WDS-IPO membership database in order to extract data about concepts and their relationships. The resulting information then needs to be stored in a triples database, published, and made discoverable. To do this, we propose development of service components that can be reused in the future:
    i. Extract triplet data from a metadata record and the existing database of WDS Members held by the WDS-IPO (in one of several supported formats), and ingest these triplet data into a triple store.
    ii. Push a meta-record to a service that does the same (this allows future on-demand updates from metadata repositories to the Knowledge Network).
    iii. Process the entities in the Knowledge Network against Linked Open Data (concept) repositories.
    iv. Implement queryable service interfaces to the Knowledge Network repository.
    v. Implement a portfolio of services based the use cases listed below.

The situations in which we envisage the Knowledge Network repository to be of value includes the following use cases:

1. WDS Membership Information

A WDS Member (or the WDS-IPO) wishes to obtain information about WDS membership. They want to

(a) Get a list and details of Members in the various WDS membership categories;
(b) Find other Members with similar functions, data, or services;
(c) Find other Members using similar technologies and/or standards.

2. Scientist Preparing a Project Proposal

A scientist wants to

(a) Find accredited repositories for their type of data;
(b) Locate service providers to help with data quality control, data processing, data publication, visualisation, data mining, and data integration;
(c) Advertise for collaborators;
(d) Identify centres/repositories using technologies, standards, and topics with which they are familiar;
(e) Locate data of interest by type, topic, theme, parameter, instrument, method of processing, author, publication centre, project name, and so on.

3. Science Funding Entity Assessment Aid

A science funding entity is assessing proposals for an international science programme/collaboration. They would like to evaluate aspects/merits of proposals in terms of

(a) Data publication/sharing—by understanding the character of the patronised networks, or extent or reach of that sharing;

(b) Data archiving—by understanding the accreditation (if any) of nominated repositories, their affiliations, or their reputation;
(c) Development of data products—by understanding synergies with existing products, or any duplication with other similar products;
(d) Novelty of the science with respect to what has gone before and what is happening now—by understanding publication history in the topic area, data produced in the topic area, or obvious gaps that need to be addressed.

4. Funding Entity Using Knowledge Network To Help Structure Calls For Proposals

A funding entity wishes to issue a call for projects when structuring a call for proposals. They might want to

(a) Understand data gaps by topic, spatial data distribution, in time; or by parameter, etc.;
(b) Channel data/products through specific available infrastructure/networks to help build/enhance them, and so need to know what exists and their capabilities.