Plinth: a basis for scalable, inter-disciplinary data usage

Authors: Phil Archer , Simon Cox , Stephen Passmore , Lesley Wyborn , Helen K Green , Simon Hodson , Laura Merson , Virginia Murray , Angelina Kraft

Since the beginning of the open data movement, many platforms and tools have been created to integrate disparate datasets and, from that integrated set, derive new insights. Two of the pilot projects within the CODATA-led data integration initiative exemplify this: The Infectious Diseases Data Observatory and Resilience.io. Such tools are part of an ecosystem around the sharing of research data that includes:

* thousands of data repositories;
* catalogues of data repositories (re3data, Core Trust Seal);
* data & metadata aggregations (DataOne being a prime example);
* Methods and tools for identifying referencing and linking datasets and scientific papers (Crossref and DataCite);
* research infrastructures and virtual research environments (EPOS, ENVRI+, EVER-EST, VRE4EIC etc ;
* existing initiatives to somehow integrate and rationalise existing vocabularies, metadata and information models and ontologies, such as FairSharing and a newly formed GODAN group on this, as well as efforts at the RDA;
* crosswalks between identifiers (Open Phacts being a prime example).

It’s a crowded space with many world-class efforts that are emerging as the de facto standard tools and methods to use.

And yet, there’s something missing. It is still not yet routine for one researcher’s dataset to be used by another, particularly cross-domain. Therefore, relationships that exist between disparate natural and social phenomena have been observed and recorded remain undiscovered for want of the tools and research time to bring them into the open.

All of the initiatives cited and many more assume to one degree or another that it will be a human who looks for, analyses, downloads and merges data. This is a well founded assumption. In both IDDO and Resilience.io, years of human effort and globally rare expertise have been spent integrating administrative and geospatial data with epidemiological, social and environmental data. Sources range from well-documented datasets that follow published standards through to countless PDF reports from which tables of statistics have been manually re-entered into a spreadsheet.

The third pilot in the CODATA-led initiative is even more manual. The task of assessing Disaster Risk Reduction steps against the Sendai Framework requires the manual discovery and collation of source reports that, for the most part, make no use of commonly agreed methodologies and terminologies, even where they exist. As a stark example, there is no commonly agreed definition of ‘a mortality’ in this context.

The Plinth proposal addresses the issue of scale. It is designed to reduce the human effort required to add more data sources as input to research environments. It follows that Plinth will make use of automation and machine learning as far as possible. Total automation is unrealistic - there’s no magic method of querying all human knowledge - but Plinth will make a start and, as its name suggests, is conceived as an infrastructure on which future projects can be built.

Plinth is not a data source but a data broker, that is, an intermediary system between the platforms created and operated by the pilots on the one hand, and data stored in repositories and elsewhere on the Web on the other. Such a system will incorporate a data model expressed as a set of ontologies and it is this model that provides the semantics that allow the clients to query disparate data sources.

The Plinth will have two primary functions: first, the ability to discover new datasets related to a given topic. Existing datasets around a specific topic will be used to train a machine learning algorithm that can then crawl the Web, particularly research data repositories, and identify new relevant datasets. The algorithm will recognise patterns in the data, as well as metadata, and so the potential exists to identify and make use of less-well documented research data that may otherwise go unnoticed by other researchers. This approach is expected to be particularly good at extending the scope of a client so that, for example, one that can initially answer questions about a specific disease in one part of the world can ask the same questions about the same disease in another area, or about a different disease in the same area.

The second primary function is a method for taking a single query that is deconstructed and executed against multiple relevant datasets to which it has online access, using whatever query language those datasets require, and the result compiled and returned as a single answer. This means that the integration step occurs at run time, not as part of an ingestion process that creates a single homogeneous data source. It depends on the identification and, where necessary the development of, ontologies and crosswalks that provide a semantic framework in which computers can ease the researcher’s task in seeking new insights.

Both of these functions have been proven elsewhere. The former as part of machine learning-based sentiment analysis used by, for example, the advertising industry to measure the effectiveness of their campaigns; the latter in, for example, Sparkall, developed at the Fraunhofer Institute.

Plinth will also offer further functions, such as automated extraction of tabular data from PDFs, the automated provision of markup based on schema.org to improve search results for humans, and more.