CRISs and persistent identifiers: how do they work together?

The scope of Current Research Information Systems (CRIS) has expanded over the years, as have the standards and technologies to support them. From being conceived mainly as internal systems for research management, the emphasis has shifted to interoperability and combining data from diverse sources. Technological developments have enabled this expansion, but a particular driver has been the trend to open science on a global scale, with transparent access to the artefacts of the research process made available for examination and reuse. The growth in repositories of publications and datasets has fuelled both the need and the opportunity for facilitating the research process through locating and making reusable these artefacts, which requires models, techniques and tools capable of handling the necessary connections at scale. (See https://ec.europa.eu/research/openscience/index.cfm for the European Commission's view on implementation of Open Science.)

The view of the CERIF standard, underpinning many CRISs, reflects this evolution: 'Today CERIF is used as a model for implementation of a standalone CRIS (but interoperation ready), as a model to define the wrapper around a legacy non-CERIF CRIS to allow homogeneous access to heterogeneous systems and as a definition of a data exchange format to create a common data warehouse from several CRIS.' (euroCRIS, 'Main features of CERIF')

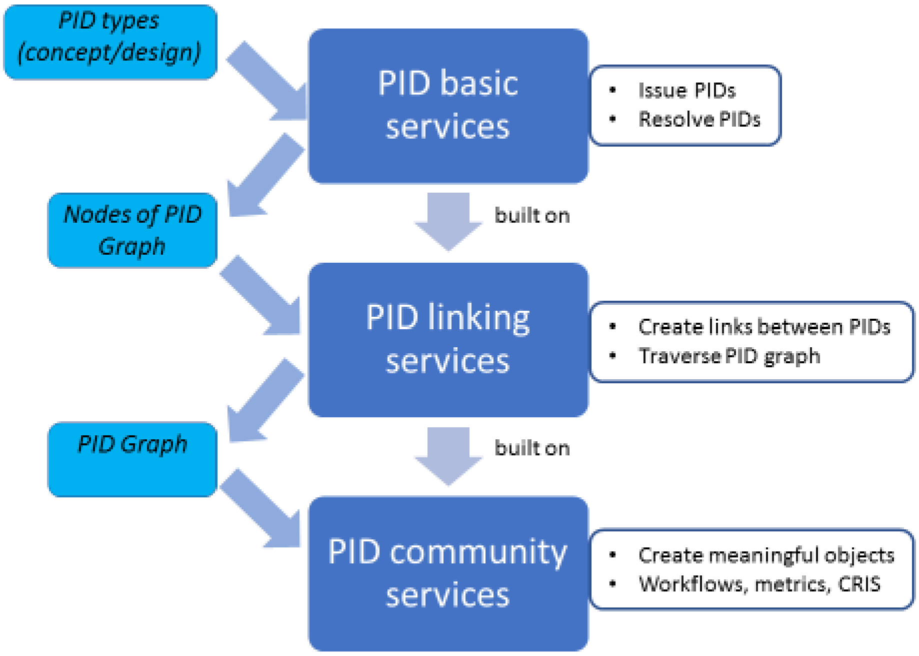
In parallel, another development has been taking place: the growth of persistent identifiers for a wide range of entities in the research domain, both in the digital and physical worlds. DOIs for publications and datasets and ORCID ids for individual researchers are now securely established, indeed indispensable parts of the research landscape, with their key features of being resolvable and having associated metadata. Taking the idea of PIDs to an extreme is the 'Global Digital Object Cloud' (RDA Data Fabric Interest Group, 'Global Digital Object Cloud (DOC) - A Guiding Vision'), putting (virtualised) digital objects centre-stage and giving PIDs a crucial role in accessing and managing them.

An earlier paper in the CRIS conference series did indeed raise this issue (Jörg et al., 2012). It acknowledged that 'System internal identifiers work well within system boundaries … but they do not scale for usage across systems. … Means to identify and consequently connect system-internal with non-internal entities are therefore needed globally, but as well within organization boundaries spanning multiple systems.' This led to the modelling in CERIF of a Federated Identifier entity.

Without necessarily adopting the thorough-going conception of the Global Digital Object Cloud, the expansion of PIDs for a diversity of objects ineluctably leads to the notion of a graph—a graph in the mathematical sense, with nodes and edges, where the nodes represent the discrete objects (digital or real-world), identified by PIDs, and the edges are the links between them, representing such connections as 'author of', 'funded by', 'cites', and 'based on [publication–dataset]'.

The Research Graph (http://researchgraph.org) provides an exemplar of this perspective. Its aim is to 'connect research publications and datasets together on the basis of co-authorship or other collaboration models such as joint funding and grants', The home page of the initiative displays impressive visualisations which make absolutely clear the transformation of the underlying data into a connected graph.

Another initiative based on the same concept is the FREYA project (http://www.project-freya.eu), which aims to construct a 'PID Graph' that 'connects and integrates PID systems, creating relationships across a network of PIDs and serving as a basis for new services'. FREYA explicitly introduces the idea of services, and envisages three levels as shown in the figure below.



It is apparent that there are points of contact between the two perspectives, CRIS-based and PID-based. They both deal in the same research-related entities, and they both represent properties of those entities—'metadata'—and links between them. A PID-centric view envisages navigating links between objects, which allows the answering of CRIS-type questions through 'services' that operate over the graph. Without setting up a false dichotomy between the two, it seems worth exploring the points of contact and the differences, and how they can work together effectively.

How then to explore the points of contact? Setting aside the somewhat philosophical question of the 'meaning' of objects and identifiers in the two approaches, there are two general aspects that can be compared: roughly speaking, 'how the data goes in' and 'how the information comes out'. Sources of the base data or metadata (about individuals, datasets, organisations, etc.) have different degrees of proximity to the database or graph, and may be liable to different extents to duplication, inconsistency or inaccuracy. Queries over databases or 'services' over connected graphs have their strengths and weaknesses.

Three generic use cases are considered: research impact assessment, individual attribution, and scientific reproducibility. They are analysed in the terms outlined above to present some conclusions about the fruitful relationship between CRIS and PID-centred systems.

References

euroCRIS. 'Main features of CERIF' [web page]. Retrieved from https://www.eurocris.org/cerif/main-features-cerif

Jörg, B., Höllrigl, T., Sicilia, M.-A. 'Entities and Identities in Research Information Systems'. CRIS 2012 Proceedings.

RDA Data Fabric Interest Group, 'Global Digital Object Cloud (DOC) - A Guiding Vision' [web page]. Retrieved from https://www.rd-alliance.org/group/data-fabric-ig/wiki/global-digital-object-cloud