Institution: Cardiff University

Unit of Assessment: 11

Title of case study: Enabling the Catalogue of Life to index the world's species

1. Summary of the impact (indicative maximum 100 words)

The loss of biodiversity is an issue of global concern. This has prompted intergovernmental aims and global campaigns, administered by organisations such as the World Wide Fund for Nature and the International Union for Conservation of Nature, to halt the rate of species extinction. A major hurdle in these initiatives was the lack of any form of definitive list of the World's species. Species data was scattered across hundreds of local databases, created and interpreted differently by many scientists. No uniform, agreed catalogue existed. However, research produced at the School of Computer Science, at Cardiff University, resolved this. The use of data modelling, constraint checking techniques, protocols and processes to amend conflicts have enabled Species2000/ITIS to produce the Catalogue of Life: www.catalogueoflife.org. This federated database is the most complete set of species data anywhere in the world, comprised of 1.4 million entries. It is accessed by approximately 30,000 users worldwide, each month, and utilised by governments across the globe for nature conservation, import control and predicting the effects of climate change. Other users include charities, specialists, scientists, publishers, students and members of the public worldwide. Therefore the categories of impact claimed are threefold - environmental, economic and impact on society, culture and creativity.

2. Underpinning research (indicative maximum 500 words)

Since 1999, members of Cardiff's School of Computer Science & Informatics have conducted basic research on the distributed data management infrastructure and associated tools for creating the Catalogue of Life, performing quality checking through conflict resolution techniques, and delivering its species data to users. Because the Catalogue is assembled from species records prepared and updated by many groups of experts around the world, the infrastructure enabling the Catalogue uses a federated approach that resolves the problems associated with unreliable heterogeneous information sources from multiple content providers. Cardiff's research has led to an infrastructure that incorporates tools for preparing the Catalogue and for maintaining its consistency, as the available data sources increase and are updated, and quality control mechanisms, in the face of the changing views of specialists on the taxonomic relationships between different organisms. This was achieved by creating a scalable architecture. Using this software the CoL has expanded every year from an initial version created as a prototype using the research, with 12 databases and 200,000 species, to its present state of 132 databases with 1.4 million species and 30,000 web users per month.

Cardiff researchers and roles over the research period: WA Gray (Prof & Principal Investigator, 1999-present), NJ Fiddian (Prof, 1999-2008), SM Embury (Lecturer, 1999-2001). AC Jones (Lecturer, 1999-2003; Senior Lecturer, 2003-present), RJ White (Lecturer, 2003-2011), A Hardisty (Manager, 2002-present), J Giddy (RA, 2002-present), ER Orme (RA, 2007-2008), N Pittas (Software Engineer, 2002-2005), H Raja (RA, 2010), I Sutherland (RA, 2000-2001), X Xu (RA, 1999-2005).

Specific contributions to the body of knowledge include:

- *Application of constraint techniques* (1999 onwards). Techniques for dealing with names whose structure and interrelationships are constrained by professional practice were developed, and constraint repair techniques were extended to address practical resolutions to conflicts discovered. Constraints pertaining to good taxonomic practice were developed in order to identify taxonomic conflicts in individual species databases and databases formed by merging multiple sources. Also developed was a method for incremental conflict resolution [3.1, 3.2]. Key

researchers: Embury, Gray, Jones.

- *Distributed architectures and protocols* (2000 onwards), including the design, construction, evaluation and deployment of a series of implementations of the Catalogue architecture. The federated approach introduced partitioned the task of maintaining a consistent classification into manageable sub-tasks. This federated structure had to be scalable to cope with the growing number of databases incorporated and the diversity of software used to maintain them. The work carried out provided an insight into interoperability with a common data model, and into the relative efficiencies of CORBA and HTTP-based infrastructures. Threats to platform independence arising from CORBA Object Request Broker incompatibilities were identified [3.3, 3.4]. Key researchers: Embury, Fiddian, Gray, Jones.
- *Techniques to deal effectively with the consequences of change* (2007 onwards). This work extended the Catalogue of Life with globally unique identifiers and explicit metadata relating the Catalogue's concepts to each other. It was demonstrated that the Catalogue of Life may be regarded as a specialised ontology, providing knowledge that is needed to support semantic interoperability in biomedical and other disciplines when dealing with species-related data [3.5, 3.6]. Key researchers: Hardisty, Jones, White.

3. References to the research (indicative maximum of six references)

3.1 **Jones AC**, **Sutherland I**, **Embury SM**, **Gray WA**, White RJ, Robinson JS, Bisby FA and Brandt SM. Techniques for effective integration, maintenance and evolution of species databases. In *Proc SSDBM 2000*, IEEE Computer Society Press, pages 3-13, 2000. http://dx.doi.org/10.1109/SSDM.2000.869774

3.2 **Embury, SM**, Brandt, SM, Robinson JS, **Sutherland I**, Bisby FA, **Gray WA**, **Jones AC** and White, RJ. Adapting integrity enforcement techniques for data reconciliation. *Information Systems*, 26, 657-689, 2001. http://dx.doi.org/10.1016/S0306-4379(01)00044-8

3.3 **Xu X**, **Jones AC**, **Pittas N**, **Gray WA**, **Fiddian NJ**, White RJ, Robinson J, Bisby FA and Brandt SM. Experiences with a hybrid implementation of a globally-distributed federated database system. In *Proc WAIM 2001 (Lecture Notes in Computer Science 2118*), Springer-Verlag, pages 212-224, 2001. http://dx.doi.org/10.1007/3-540-47714-4_20

3.4 **Xu X**, **Jones AC**, **Gray WA**, **Fiddian NJ**, White RJ and Bisby FA. Design and performance evaluation of a web-based multi-tier federated system for a catalogue of life. In *Proc 4th international workshop on web information and data management (WIDM 2002),* ACM Press, 104-107, 2002. http://dx.doi.org/10.1145/584931.584954

3.5 **Jones AC**, **White RJ**, **Giddy J**, **Hardisty A** and **Raja H**. Evolution of the Catalogue of Life Architecture, In *Knowledge-Based and Intelligent Information and Engineering Systems (Proc KES 2012, Lecture Notes in Computer Science Volume 6279)*, Springer-Verlag, pages 485-496, 2010. http://dx.doi.org/10.1007/978-3-642-15384-6_52

3.6 **Jones AC**, **White RJ** and **Orme ER**. Identifying and relating biological concepts in the Catalogue of Life. *Journal of Biomedical Semantics* 2(7), 2011. http://dx.doi.org/10.1186/2041-1480-2-7

4. Details of the impact (indicative maximum 750 words)

The Catalogue of Life is endorsed by the international UN Convention on Biodiversity (CBD). Funding for its establishment and continuing operation was provided by multiple EU framework projects [5.10] and the Global Biodiversity Information Facility, GBIF ($565,000) [5.3]. It is the world's most authoritative source of peer-reviewed information about the names (Latin scientific names and common names) of the world's species of plants, animals, fungi and micro-organisms. It currently holds entries for more than 1.4 million (out of an estimated 1.9 million) species. The existence of the catalogue, stemming from an early prototype in 1997 to its current state in 2013, is

a direct consequence of Cardiff University's research. Dr Peter H Schalk, ETI BioInformatics and current Chairman of the Board of Directors of Species 2000 (a not-for-profit organisation set up for the delivery of the catalogue) commented that "the extended coverage of the CoL (from 600,000 in the late 90s to 1,400,000 species now) was made possible because of software developments which underpin the complicated data management process. Cardiff played a crucial role in advising how to approach the problems of up-scaling (from less than 30 providing databases to over 100) enhancing efficiency (one update per year to 4-6 updates at present) and professionalisation (developing proper software tools streamlining the CoL production process). The current CoL data management architecture is largely based on innovations and prototype developments carried out by Cardiff. … without the work carried out by the Cardiff team the CoL would not have developed into the global resource it is today" [5.1].

Since 2008 the catalogue has had 30% new users. In 2011 unique visitors to the website generated 70-90 million page hits. In 2012, there were 3,777,000 hits from people in 212 countries, the top five being United States – 60589, France – 38649, UK – 22273, Spain – 18447 and Germany – 15341. Each year, in addition to web usage, 3,500 physical CD/DVD copies of the Catalogue are distributed to 80 countries [5.2].

Specific examples of usage of the catalogue during the REF period, encompassing environmental, societal and economic benefits, are as follows (note that access data is given for 2012, the most recent year for which complete data and breakdowns are available):

- Use in the preparation of the Red List to check the information about species being added to the endangered species list, to identify all of the synonyms. This resulted, for example, in 80,000 hits in 2012 [5.1].
- Use as a "taxonomic backbone" (via web services and data download) to a variety of international scientific data sharing initiatives [5.4; see "Minutes of Evidence" para 40 (page 62)]. Two significant users of CoL are GenBank (a database administered by NCBI, a centre under the US Government's National Institute of Health) and the European Bioinformatics Institute (EBI) which use the Catalogue as a source of authoritative species and organism data in DNA sequence searches. In 2012 hits from these sources reached 63,000 [5.1].
- Use in underpinning the Encyclopedia of Life, by checking that names are valid before being added to the online encyclopedia, led to 20,000 accesses to CoL in 2012. Other organisations that use the CoL in a similar respect are the International Union for the Conservation of Nature (IUCN: www.iucn.org), and the Consortium for the Barcode of Life (CBOL: www.barcoding.si.edu) – each, in turn, have substantial worldwide usage [5.1, 5.6].
- Multi-national trading companies like IKEA need documentation systems for the provenance of their raw materials, arising from recently enacted American law – the Lacey Act (2009) www.eia-global.org/lacey. CoL was the basis of IKEA's first set of data for the first properly declared list of furniture species and the data is used in 800 IKEA factories around world for preparation of data for importation to the USA. IKEA paid Species 2000 $50,000 for this access [5.7].
- Academic publishers including Taylor and Francis and Reed Elsevier use CoL content in their e-library products. CoL provides them with authoritative sets of categorised controlled terms and phrases for species and organisms for search and browsing. Annual contracts exist with these companies which have brought in £31,000 since 2008 [5.8].
- Use in Europe to clean up the Natural History Collections in Museums to identify naming problems in collections and correct them - 260,000 zoological objects have been checked in the first three months of 2013 [5.9].
- Use by BGCI (which organises information on plants conserved in botanic gardens around the world), by the Institute of Zoology in London (which works on conservation status analyses

using national Red Lists and the Sampled Red List Index), and by European partners in the Biodiversity Heritage Library (BHL) project [5.1].

In sum the Catalogue of Life is regarded as a highly valuable resource, used on a global basis, for a variety of purposes. The initial and ongoing availability of the CoL is due to Cardiff University's research.

5. Sources to corroborate the impact (indicative maximum of 10 references)

5.1 Chairman of the Board of Directors of Species 2000. *Corroborates (1) the fact that the research has directly led to the availability of the CoL and (2) the impact derived from several organisations that use the catalogue.*

5.2 i4Life Dissemination Report, April 2013 – includes aggregated CoL web access statistics for 2012. *Corroborates the web access data given in Section 4.* [Available on request from HEI; see pages 28 onwards]

5.3 Executive Secretary of GBIF. *Corroborates (1) the use of the CoL by GBIF, (2) the role that Cardiff University's research played in the availability of the catalogue, and (3) the amount of funding provided.*

5.4 European Nucleotide Archive Team Leader. *Corroborates the use of the CoL by EMBL-EBI.*

5.5 http://www.parliament.the-stationery-office.co.uk/pa/ld200708/ldselect/ldsctech/162/162.pdf *External report from the House of Lords Systematics & Taxonomy inquiry corroborating the need for the Catalogue of Life to underpin biodiversity activity.* [Available on request from HEI]

5.6 Memorandum of Understanding between the EoL and the CoL. *Corroborates the use of the CoL by EoL.* [Available on request from HEI]

5.7 Bank statements. *Corroborates that IKEA paid to utilize the CoL*. [Available on request from HEI]

5.8 Publisher, Chemistry and Life Sciences, Taylor and Francis. *Corroborates the use of the CoL by Taylor and Francis.*

5.9 Information Manager, Naturalis. *Corroborates the use of the CoL by Naturalis.*

5.10 Catalogue of Life Funding & Support. http://www.catalogueoflife.org/colwebsite/content/contributors#5 [Saved as PDF 31/10/13; available on request from HEI] *Corroborates CoL funding from multiple EU framework projects.*