Institution:
Edinburgh Napier University

Unit of Assessment:
11

Title of case study: Standards for Taxonomic Classification of Biodiversity Data

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

At Edinburgh Napier University Professor Kennedy's research on modelling processes and results of biological classification has had, and continues to have, a major impact on the infrastructure for storing and exchanging biodiversity data worldwide. It led to the Taxonomic Concept Schema (TCS), a biodiversity data standard ratified by the International Biodiversity Standards Group (TDWG), now the basis of systems worldwide for referencing biodiversity data, including the Global Biodiversity Information Facility and the International Plant Names Index. The TCS fed into the design of the Darwin Core standard subsequently ratified by TDWG, and now the exchange format for data in the major biodiversity infrastructures globally.

2. Underpinning research (indicative maximum 500 words)

A biological classification provides a means of identifying, categorising and referring to organisms. However, the complexity of the living world, and variety of techniques for surveying it, means that the same organism may be classified differently according to taxonomic opinions, and known under several alternative names. Even for well-studied large mammals there is disagreement over their correct classification and names of species, e.g. since 1812 there have been nine different classifications of Gorilla, each containing different sub-species with different circumscriptions.

Prior to 2000, taxonomic databases could model only one classification where there should be many, which obscured the uncertainty inherent in taxonomy [Paper-3]. Working with the Royal Botanic Garden, Edinburgh [Grant-6], in 1998, Kennedy developed a model and database system (Prometheus) to accommodate multiple overlapping hierarchies, accurately representing the process of biological taxonomy [Paper-1]. This was enhanced with visualisations to help botanists understand the differences and similarities in how taxonomists classify specimens [Paper-2].

International recognition of the value of Prometheus led to collaboration with seven US Institutions on the Scientific Environment for Ecological Knowledge (SEEK) project [Grant-2, 2002 – 2008]. Kennedy was a key member the Taxonomic Working Group, modelling taxonomic concepts for the SEEK taxon database. Kennedy then applied her research to support ecologists, who use species names resulting from taxonomy to identify and record species' occurrences for biodiversity analysis. SEEK aimed to integrate diverse datasets on species' occurrences across wide geo-temporal ranges, complicated by the uncertainty in name usage and variety of data formats to record species. Therefore a standard exchange format was required.

In 2005 the International Biodiversity Information Standards Group (TDWG) invited Kennedy to lead the development of an international data standard for describing taxonomic names and concepts [Grant-4, Grant-5]. This involved collaboration with major biodiversity groups to understand their perspectives on taxonomy, and to resolve taxon ambiguity issues that plagued biodiversity data products. Kennedy developed the Taxonomic Concept Schema (TCS) [Paper-4], separating names from the underlying concepts, which galvanised the community by supporting the exchange of both semantically poor legacy and semantically rich data sets. It was used as the format for novel visualisation techniques for exploring multiple classifications [Paper-2, Paper-5] helping educate the community. TDWG ratified the TCS XML schema as an international standard in 2005. It has since served as a concrete model for interoperability among data systems.

As Research Theme Leader at the e-Science Institute, Edinburgh University in 2005-6 Kennedy hosted several workshops to drive forward the development of standards for integrating scientific data. This work included development of an ontology for TDWG based on TCS.TCS concepts were then incorporated into Darwin Core (DwC) a simple file transfer format used by museum collections. This was subsequently ratified as a standard by TDWG in 2009 [Evidence-7]. The Darwin Core Archive is the file format used for the exchange of data based on terms specified in DwC.

Kennedy's expertise in TCS development and species names and concepts contributed to the publication of the *Minimum Information about a Genome Sequence Specification* (MIGS). This has stimulated significant research in the genomics area as evidenced by 442 citations of [Paper-6].

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

The following references are the key journal publications arising from the work on taxonomy and visualisation. Each is an internationally significant publication in its field.

Paper-1: Pullan, M.R., Watson, M., **Kennedy, J**., Raguenaud, C., Hyam, R. (2000). The Prometheus Taxonomic Model: a practical approach to representing multiple classification. *Taxon, 49,* 55-75.

Paper-2: Graham, M., **Kennedy, J**. (2005). Extending taxonomic visualisation to incorporate synonymy and structural markers. *Information Visualization, 4*(3), 206-223.

Paper-3: **Kennedy, J.**, Kukla, R., Paterson, T. (2005). Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration. In: Ludaescher, B., Raschid, L. (Eds.) *Data integration in the life sciences, 3615.* (pp. 80-95). Berlin Heidelberg: Springer-Verlag.

Paper-4: **Kennedy, J**., Hyam, R., Kukla, R., Paterson, T. (2006). A standard data model representation for taxonomic information. *OMICS: A Journal of Integrative Biology, 10*(2), 220-230.

Paper-5: Graham, M., **Kennedy, J**. (2007). Exploring multiple trees through DAG representations. *IEEE Transactions on Visualization and Computer Graphics, 13*(6), 1294-1301.

Paper-6: Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M., Angiuoli, S., Ashburner, M., Axelrod, N., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., Vos, P., dePamphilis, C., Edwards, R., Faruque, N., Feldman, R., Gilbert, J., Gilna, P., Glöckner, F., Goldstein, P., Guralnick, R., Haft, D., Hancock, D., Hermjakob, H., Hertz-Fowler, C., Hugenholtz, P., Joint, I., Kagan, L., Kane, M., **Kennedy, J**., Kowalchuk, G., Kottmann, R., Kolker, E., Kravitz, S., Kyrpides, N., Leebens-Mack, J., Lewis, S., Li, K., Lister, A., Lord, P., Maltsev, N., Markowitz, V., Martiny, J., Methe, B., Mizrachi, I., Moxon, R., Nelson, K., Parkhill, J., Proctor, L., White, O., Sansone, S., Spiers, A., Stevens, R., Swift, P., Taylor, C., Tateno, Y., Tett, A., Turner, S., Ussery, D., Vaughan, B., Ward, N., Whetzel, T., Gil, I., Wilson, G., Wipat, A. (2008). The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology, 26*, 541-547.

Funding for Research

Grant-1. Kennedy, J. Visual Exploration of Species-referenced Repositories (VESpeR) (2012-14), funded by EPSRC/BBSRC bioinformatics initiative, BBSRC, £106,904.

Grant-2. Kennedy, J. SEEK: Science Environment for Ecological Knowledge (2002-2008) $350,000 (of $12.5m) Funded by National Science Foundation, USA

Grant-3. Kennedy, J. TaxVis: Visualisation Tools for Integrating Large Alternative Linnaean Taxonomies (2006-2008) Funded by EPSRC £140,000

Grant-4. Kennedy, J. TDWG Core Ontology (2006) TDWG/GBIF (Betty & Gordon Moore Foundation) $65,000

Grant-5. Kennedy, J. TDWG Taxon Concept Transfer Schema (2004-2005) TDWG/GBIF $60,000

Kennedy, J. Prometheus (1998-2000) funded by BBSRC £80,000

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

Following the ratification of the TCS as an international standard in 2005, numerous governmental and NGO agencies have benefitted from the Taxonomic Concept Schema (TCS) as developed by Kennedy. Equally, the Darwin Core (DwC) standard, which incorporates concepts from the TCS, has significant impact. The impact of the research during the assessment period is most visible in three broad communities.

1. Biodiversity Science Organisations:

The adoption of TCS and follow-on standards such as DwC, has significantly increased the global availability of biodiversity data. These standards now define best practice for a number of professional bodies, and as such have significant impact on *practitioners and professional services* in the field. Adoption of the technology by NGOs and other public sector organisations also impacts *society, culture and creativity*. This enables scientists in academic, governmental and industry to access and share biodiversity information to help understand the issues in global biodiversity, and thus influences both research and policy debates. Example biodiversity organisations include:

- In 2009 the **Global Biodiversity Information Facility** (GBIF), established by governments to encourage open access to biodiversity data, supported TCS in its Integrated Publishing Toolkit. This encouraged the uptake of TCS around the world [Evidence-1]. DwC-A, informed by TCS, became the preferred format for publishing data to the GBIF network in 2012 and has been used to mobilise the vast majority of specimen occurrence and observational records within the GBIF network [Evidence-2]. GBIF currently has 416,242,316 indexed records, 10,140 datasets, 562 publishers, 53 countries signed up, GBIF mediated data used/cited in over 8000 papers.
- In 2012, the **Royal Botanic Gardens Kew** [Evidence-3] adopted TCS as its internal transfer format. The attraction of TCS to Kew is that it supports the separation of nomenclature and taxonomy, as advocated by Kew's Science and Horticulture Systems (SHS) Project. SHS is one of the component projects of Kew's IT and DM Strategy Programme, the mission of which is to modernise, integrate and streamline all the information systems that address science and horticulture information systems, as well as manage several millions of records of data, across all aspects of Kew's activity. **The International Plant Names Index** (IPNI) at Kew, a collaboration between The Royal Botanical Gardens Kew, Harvard University and the Australian National Herbarium, uses TCS. Its impact globally is clear: TCS serves information on 1,624,845 name citations, 43,011 authors and over 17,066 publications.
- **ZooBank**, which is the official registry of Zoological Nomenclature, adopted the use of taxon-name-usage in 2009. This is an example of species concepts based on TCS [Evidence-4]. As with GBIF and INPI, this system has enormous influence on biodiversity research. It contains details on 106,918 Nomenclatural Acts, 41,880 Publications and 22,150 Authors.
- The **Catalogue of Life** is the most comprehensive and authoritative global index of species currently available. It consists of a single integrated species checklist and taxonomic hierarchy. The TCS was adopted for the Catalogue of Life in order to express information about concepts and their relationships through their life science identifiers in 2008. The Catalogue now holds essential information on the names, relationships and distributions of over 1.4 million species. [Evidence-5]

2. General Public

The use of TCS has also impacted society by stimulating public interest and discourse in science. For example, the **Encyclopedia of Life** (EOL) gathers, generates, and shares knowledge in an open, freely accessible and trusted digital resource in order to achieve a vision of global access to knowledge about life on earth. The EOL harvests content prepared according to the GBIF Darwin Core Archive (DwC-A) [Evidence-6], the format derived in part from TCS [Evidence-7]. This

resource currently has 1,364,055 pages, 69,552 members, 5,825 collections and 203 communities.

3. Individual Scientists outside academia

The research has influenced the scientific practice of modern taxonomists both directly and through tool support

- **Scratchpads** is an online virtual research environment for biodiversity scientists that facilitates the free sharing of data and the creation of research networks. It uses TCS as an information exchange format [Evidence-8]. The uptake of this technology is vast. There are currently 572 Scratchpads, used by 6,851 active users covering 76,387 taxa in 509,237 pages.

Projects such as the **Global Names Usage Bank** and the Taxonomic Name Resolution Service have built upon TCS concepts to produce the most valuable taxonomic concept data available. This will significantly ease the process of data integration that has proved so difficult in biodiversity projects to date. Concepts are now mandated in the US National Vegetation Classification [Evidence-9]. The National Centre for Ecological Analysis and Synthesis, USA utilises services such as TNRS in many of its informatics projects (e.g., the **Botanical Information and Ecology Network**, BIEN) and recommends its use to biodiversity scientists worldwide for synthesis projects [Evidence-10].

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

Evidence-1. http://www.e-biosphere09.org/assets/files/e-Biosphere%20Abstracts%20Volume%20-%20FINAL.pdf, page 116.

Evidence-2. http://www.gbif.org/informatics/standards-and-tools/publishing-data/data-standards/darwin-core-archives/ Contact: Information Architect, GBIF, Copenhagen

Evidence-3. http://www.ipni.org/stats.html Contact: Head of Nomenclature and Taxonomy (Biodiversity Informatics), Kew Gardens, London

Evidence-4. http://iczn.org/files/BZN%2066%284%29_Unifying_nomenclature.pdf Contact: Associate Zoologist, Bishop Museum, Hawaii

Evidence-5. Jones AC, White RJ, Orme ER. Identifying and relating biological concepts in the Catalogue of Life., J Biomed Semantics. 2011 Oct 17;2(1):7. doi: 10.1186/2041-1480-2-7.

Evidence-6. http://eol.org/info/cp_archives

Evidence-7. http://rs.tdwg.org/dwc/ Contact: Museum of Vertebrate Zoology, Berkeley, California

Evidence-8. http://www.isgtw.org/feature/vibrant-time-biodiversity Contact: Natural History Museum, London, dmr@nomencurator.org

Evidence-9. http://www.fgdc.gov/standards/projects/FGDC-standards-projects/vegetation

Evidence-10. Contact: National Centre for Ecological Analysis and Synthesis, Santa Barbara, USA. http://www.globalnames.org/GNUB); http://bien.nceas.ucsb.edu/; http://tnrs.iplantcollaborative.org/. NB. This is not a university department, but a governmental group based at ucsb