Biodiversity Information Standards
TDWG

Standards for the Exchange of Biodiversity Data

http://www.tdwg.org

Natural Collections Description (NCD)

A data standard for exchanging data describing natural history collections

Neil Thomson (Natural History Museum, London), Roger Hyam (Royal Botanic Garden Edinburgh), Constance Rinaldo (Harvard University), Carol Butler (Smithsonian Institution), Doug Holland (Missouri Botanical Gardens), Barbara Mathé (American Museum of Natural History), Günter Waibel (RLG Programs, OCLC), Wouter Addink (ETI Bioinformatics), Ruud Altenburg (ETI), Markus Döring (Berlin Botanic Garden)

1. Summary

Note: *This is a non-normative companion document, which provides some background to the aims and uses of the proposed standard. The normative document has been separated and may be found on the TDWG Website at* http://www.tdwg.org/activities/ncd/

Natural Collections Description (**NCD**)[1] is a data standard for describing collections of natural history materials at the collection level; one NCD record describes one entire collection.

Collection descriptions are electronic records that document the holdings of an organisation as groups of items, which complement the more traditional item-level records such as are produced for a single specimen or a library book. NCD is tailored to natural history. It lies between general resource discovery standards such as Dublin Core (**DC**) and rich collection description standards such as the Encoded Archival Description (**EAD**). It is possible to extract a Dublin Core record from an NCD record for use with general resource discovery systems, or to use an NCD record as the seed for a richer collection description, like an EAD record.

The NCD standard covers all types of natural history collections, such as specimens, original artwork, archives, observations, library materials, datasets, photographs or mixed collections such as those that result from expeditions and voyages of discovery.

NCD primarily holds information about collections of objects, but can also be used to describe organisations (collections of collections) and networks (collections of organisations). There are many existing sources of information about biodiversity organisations, but they are scattered and in different formats.

This document accompanies the normative part of the NCD standard. The standard consists of the series of class and property definitions and is presented in a separate document. These definitions are identified by unique TDWG Uniform Resource Identifiers (URI). It assumes that copies of the definitions will be hosted at the

[1] A glossary of acronyms is provided at Appendix 2. First usage of an acronym is in bold font.

specified URIs given in the normative form document. The URL stem is **http://rs.tdwg.org/ontology/voc/**

The standard also contains recommendations on the use of Dublin Core and vCard properties. This avoids duplicating established vocabularies and facilitates interpretation of NCD documents by non NCD aware applications.

It is expected that NCD will develop further as experience is gained in the projects that are making use of it, particularly in the addition of terms to the pick-lists. It has reached a sufficiently mature state that applications may be based on it. Following approval from the TDWG appraisers it will be designated as NCD version 1.0 and made widely known to the biodiversity informatics community.

The NCD standard is the culmination of work on collections descriptions carried out for the European Union Framework VI programme SYNTHESYS and the work performed by the Anglo-American group Resources Available in Natural Sciences (**RAVNS**), which operates under the auspices of RLG Programs, OCLC.

The normative documentation includes an example record and, for those that are developing applications based on NCD a column provides suggested cardinality – that is, for fields that should be considered mandatory (**M**) or may have more than one value – that is, repeatable (**R**) or indicate text in the local language (**L**).

To ensure that the barriers to usage are as low as possible, only four properties of the Collection class are considered to be mandatory:

1. Author of the record
2. Date of record creation
3. Collection name
4. Collection description

An NCD Toolkit has been developed by ETI Bioinformatics in Amsterdam (http://www.eti.uva.nl/) with the aid of funding from GBIF and is available for download at https://sourceforge.net/projects/ncdtoolkit/.

Version 1.0 of the NCD Toolkit is based on NCD v0.8. It is a cross-platform database which enables natural history organisations to record data about their own collections.

2. Motivation and Rationale

Many valuable collections exist that have no information stored in databases, nor do they have a web presence. Such collections are easily overlooked by researchers, so a brief descriptive record as defined by the NCD standard can act as the “business card” for a collection, providing enough information to identify and locate it.

The standard enables the aggregation of collections descriptions from many sources and facilitates resource discovery, including establishing relationships among collections in several locations. NCD records can also be used as an aid for collections management processes, allowing an institution to take a step back and see which collections are most in need of conservation or would benefit from a higher priority for item-level cataloguing.

The standard was developed by the TDWG NCD Interest Group to fit with the suite of data standards being developed on behalf of the Global Biodiversity Information Facility (GBIF) by Biodiversity Information Standards (TDWG).

Diagram 1 gives a very simplified view of the use of these and other standards to exchange information between some of the stakeholders in biodiversity informatics.

The relationships between some of the major components are represented by reading down the left-hand column, which also shows an organisation or project that is developing each. Data interchange standards enable the flow of data between providers, users and communities and one of each is shown bridging the first and second columns, with a key to the acronyms provided at the foot of the diagram. Note that this is a two-way flow. Finally, the influence of name servers and generic data such as globally unique identifiers (**GUID**), dates and geospatial coordinates in providing consistency is shown in the right-hand column with some of the driver organisations that provide guidance on good practice.

One use for NCD can be seen within this overall picture, providing information about the collections that hold specimens. A project based on NCD (which cannot be named in this document) aims to provide a GUID for all such collections throughout the world so that researchers may unambiguously specify the source of their material and contact information for those wishing to visit collections. NCD provides a mechanism for exchanging and aggregating information about collections in a standard format so that applications based on it may share data.

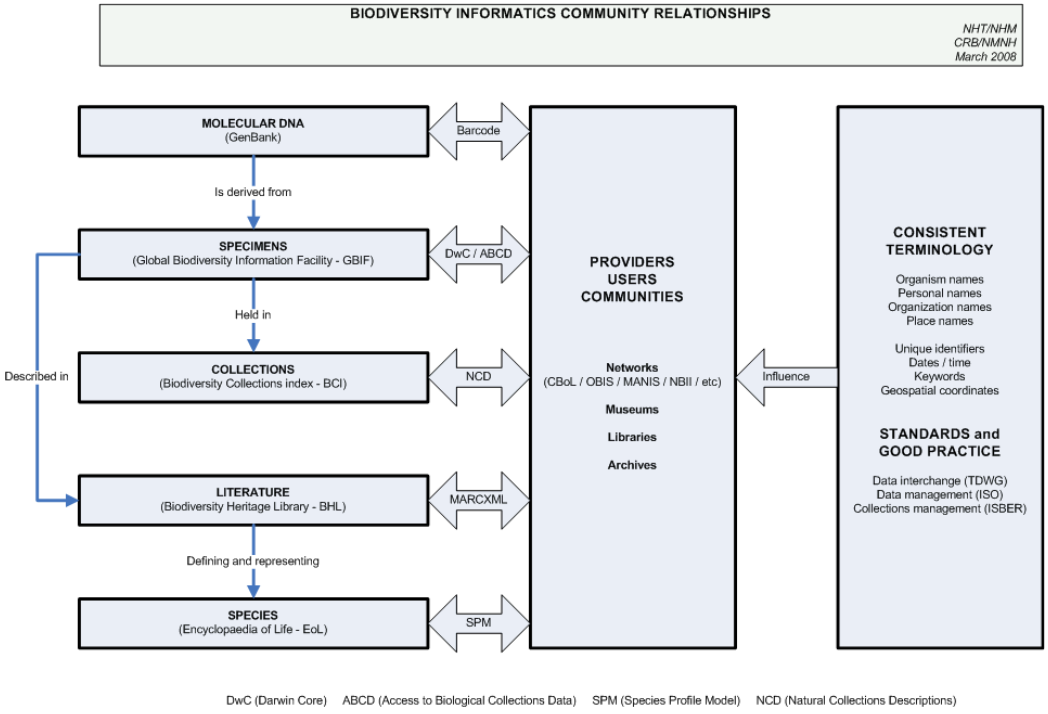

Figure 1: Simplified view of biodiversity community relationships and standards

Funding has been provided by:
- The European Commission
- Gordon & Betty Moore Foundation
- Natural History Museum, London
- RLG Programs, OCLC
- Smithsonian Institution

Grateful thanks are offered to all of these.

3. Description

3.1 Collection descriptions

Collection descriptions are electronic records that document the holdings of organisations as groups of items. Such descriptions complement the more traditional item-level records describing a single specimen or a library book. Each collection record describes one entire collection, including narrative information on the collection itself, its extent and purpose, conditions of access and use along with who to contact for more information.

A collection may be loosely defined as any group of things that have something in common. That "something in common" can be defined by the basic questions that users ask when accessing collections – who, what, where and when.

Examples of collections include:
- items that were collected or made by a particular person
- items that have the same format, such as art on paper
- items that came from the same place
- specimens that belong to the same taxonomic group
- materials collected on a voyage of discovery

In natural history museums, for example, researchers are most familiar with the collections of specimens, the library and the archives but the exhibitions, paintings, sculptures and learning materials are also collections.

Digital collections include images, video, datasets and databases (which are collections of item-level records) and the thematic sections of web sites. Noting the formats and media used to store digital data will be of value in digital sustainability planning, so that the process of migrating data from imminently obsolescent formats may be effectively managed. This will probably be carried out in conjunction with tools that are being developed by the digital sustainability community.

Collections of natural history material can be large. Consequently, detailed item-level descriptions can take a long time to complete. Collection-level records can ensure that knowledge about the richness of collections can be revealed more rapidly. Relating collections that are in museums, libraries, archives or other organisations (cross-domain resources) is a priority for many governments and by adopting the same description standard for collections in each domain, it becomes possible to search across all collections, regardless of management domain or location.

Some organisations divide collections between departments for curatorial purposes. Researchers would need to contact each department individually to assess the complete collection. Similarly, some collections have been dispersed throughout several organisations or even across several countries. These collections may be reunited in a virtual sense, using collection descriptions for each component.

A collection description record can be created for a collection whether the items in that collection have their own records in a database, or not. Where a database containing item-level details exists, a link can be provided to that database for those that need that level of detail. If the collection does not have an item-level database, producing a collection description reduces the chances of that collection being overlooked by researchers using the Web for resource discovery. Collections cannot be protected if they are not known to exist.

Collection descriptions provide a broad perspective and such records can serve a variety of additional purposes for organisations:

- A collections inventory is helpful in protecting against both loss of data and loss of collections and thus serves as a form of audit control and security against unwarranted disposal.

- They can help with the assessment of the strengths and gaps in the organisation as a whole, so that finding collaboration partners that have either the same or complementary strengths is simplified.

- They can help to identify which areas should be a priority for development in strategic plans and to establish priorities for item-level cataloguing. For conservation assessment, the McGinley scale is recommended, details of which can be found at:

  McGinley, R. J. 1993. Where's the Management in Collections Management? Planning for Improved Care, Greater Use and Growth of Collections. *In:* Rose, C. L., et al. (eds.). International Symposium and First World Congress on the preservation and conservation of Natural History Collections 3. Communidad de Madrid Consejeria de Educacion y Culturea and Direccion General de Bellas Artes y Archives Ministerio de C, Madrid. Pages 309-338.

- Collection descriptions can serve to prevent loss of data that is in a physical form or electronic data in a format or medium that is nearing technological obsolescence. Creating collection descriptions for datasets that includes format information will help to act as an early warning so that data can be migrated to a more current format. Such data then becomes part of a digital sustainability programme, rather than a digital archaeology project.

- Collection description records act as a convenient place to store information volunteered by collections managers or visitors, which may otherwise be lost on their departure.

Records can be created *de novo* or from existing resources, such as published finding aids. There are many of these, but they are all in different formats, mainly on paper and cannot easily be searched. Once collection level data exists it can be used for internal projects such as exhibition labels or for external initiatives such as the merging of data from several sources to provide regional coverage of biodiversity collections.

3.2 NCD records

An NCD record consists minimally of the 4 mandatory fields (Author, Record created date, Collection name and Description) so that it is easy to set up holding records that may be filled out when resources allow. It is suggested that each record will be serialized in the Resource Description Framework (**RDF**) and its Identifier will be a resolvable Life Sciences Identifier (**LSID**) or Uniform Resource Locator (**URL**) to that

RDF file but the standard does not mandate the use of RDF (see *Implementation and Compliance* below). All other fields are considered to be optional, but of course the more information that can be provided about a collection the more useful the record will be.

The normative documentation gives the labels, Uniform Resource Identifiers (**URI**s) and definitions for each NCD class and property. Also provided are the tables of consistent terms for use in pick-lists and an example record.

The standard caters for collections of any type of material, physical or digital and either private or corporate ownership. It may be important to distinguish between physical collections and derived collections. An example of a derived collection record is one that has been produced as the result of a query on a collections management database, such as “all the items from Australia”. This contains useful information that the institution may wish to keep, but could cause inaccurate totals if included in a count of collections held at the institution, since it does not exist as a discrete collection.

Records include information about who created the record and when, or the source of the records if they have been harvested from elsewhere. If a record is subsequently edited then the editor and date of editing may be recorded. NCD only directly addresses the most recent edit, but an edit history could be built up using the <Notes> memo field.

Many of the fields may be repeated, either to accommodate multiple entries, such as the <Associated person> property in the example, or because the entry is in more than one language. Eight of the fields have English-language controlled terms associated with them, to aid searching and sorting.

Other fields may draw terms from existing authorities and it is recommended that an indication is given of the source of those terms along with, if possible, the identifier for the authority record for the term within that source. For an example, see the <Place name coverage> property in the example record, which gives the Getty Thesaurus of Geographic Names (**TGN**) identifiers for several of the place names entered. This service may be used from http://www.getty.edu/research/conducting_research/vocabularies/tgn/

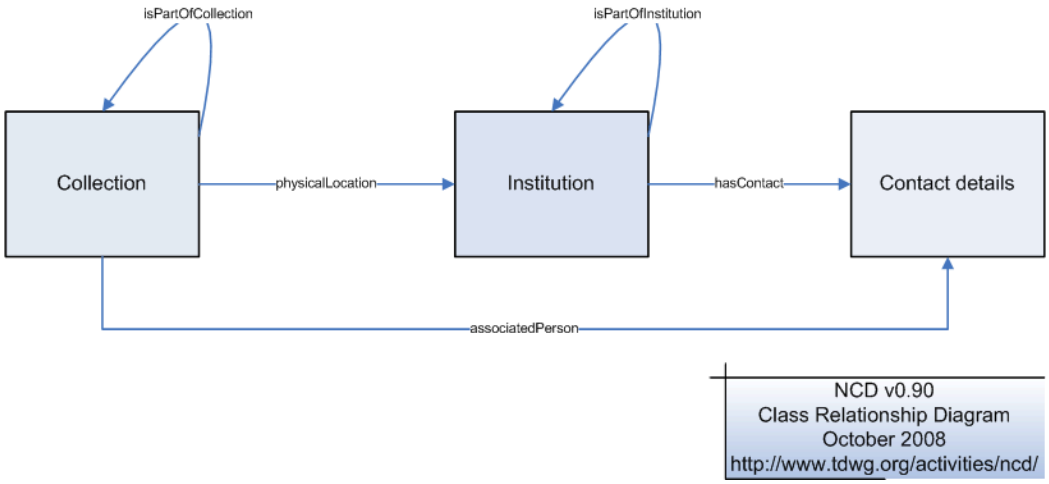

Figure 2 NCD Class Relationship Diagram

Each collection will be associated with one or more persons, through the <Associated person> property, or with an institution, which will typically be the owner and/or location of the collection. The vCard standard (http://www.w3.org/TR/vcard-rdf) has been adopted and supplemented for use in recording details about persons and institutions, since one of the main purposes for NCD records will be to find out who to contact for more information about consulting the collection.

An institution may be considered as a "collection of collections" and so has its own description property, along with a property for recording the various acronyms and codes by which it may be known. Similarly a network, such as BioCASE or the European Distributed Institute of Taxonomy **EDIT** (http://www.e-taxonomy.eu/) may be considered as a "collection of institutions".

Collections may be related to their parent collection or institution and institutions may be related to their parent institution or network so that it is possible to build hierarchies. In general, it is easier to relate upwards to a parent than downwards to children. The latter may be achieved by requesting all records that have *this* identifier in their Parent collection identifier field.

4. Implementation and Compliance

In a similar spirit to the Dublin Core metadata initiative (**DCMI**), NCD is defined in as a technology-neutral way as possible. It provides natural language definitions of classes, properties and instances that are identified by URIs and it makes recommendations on the use and content of properties from other vocabularies (Dublin Core and vCard).

The URIs defined here may be used across a number of technologies, such as namespaces in XML Schema validated documents and column headings in tab delimited text files.

This approach facilitates:

- Embedding of NCD data within other standards such as descriptions of specimens or literature.
- The extension of NCD records with other data types such as geospatial attributes.
- Cross walking between technologies such as a Comma Separated Value file, an RDF graph, an XML document and a JSON object.

The weakness of this approach is that this standard itself does not provide an off-the-shelf, self validating exchange format. The strength is that multiple such exchange formats meeting different requirements can be defined and this standard allows mapping between them.

The RDF files of the latest version of NCD may be found at: http://rs.tdwg.org/ontology/voc/ (Note: Use **View** | **Source** in a Web browser to see the actual RDF).

To implement this standard, consult the NCD Toolkit User Guide. The NCD Toolkit was developed by ETI in Amsterdam and based on NCD v0.8. Individuals and institutions that would like to start managing their collection-level records in NCD are encouraged to make use of the Toolkit, which may be downloaded from Sourceforge at the URL provided below.

The Toolkit allows the export of data in NCD format so that records may be aggregated in to regional or national systems, or into the global Biodiversity Collections Index.

5. Further Information

- NCD Website:
  http://www.tdwg.org/activities/ncd/
- To join the mailing list:
  http://lists.tdwg.org/mailman/listinfo/tdwg-ncd
- Discussion and documents:
  http://wiki.tdwg.org/twiki/bin/view/NCD/WebHome
- NCD Toolkit:
  https://sourceforge.net/projects/ncdtoolkit/

Appendix 1: NCD Development Background

The genesis of the NCD data standard can be traced back to 1999 at the Natural History Museum (**NHM**) in London, UK. At that time, a new building was being created that would allow the visiting public to view something of the extent of the working collections of specimens in addition to those highlight specimens that are normally on view in the exhibitions areas.

This building was to be called the Darwin Centre and it was intended to make the first class scientific research that is carried out at the Museum more visible and better known. The key phrase was “unprecedented access to the collections”, but there was no overall understanding of what collections existed in the Museum to which to give this unprecedented access.

Funding was provided to create collection-level descriptions (**CLD**s) for all the Museum’s collections, whether they were of specimens in the science departments; original artwork depicting those specimens held in the Library; expedition field guides held in the Archives; sculptures scattered through the galleries; learning materials or some other collection. The ability to link related, but separately managed, collections was a bonus.

Over 1,000 records were created over the next couple of years and two things became clear. One was that this was just the tip of a very large iceberg and the other was that the data standard being used was too rich for the purpose. The project made use of the archival standard Encoded Archival Description (**EAD**) since this was specifically designed to describe collections and was already in use in the Museum Archives.

Thoughts turned to creating a simplified metadata standard that would serve to record guide-book style information about collections which could be amplified at a later stage, if necessary.

About this time there were two relevant external developments. One was a rising interest in CLDs, mainly driven by the Research Support Libraries Programme (**RSLP**) in the UK. This was developing a generalised simple collection description standard which would eventually become one of the Dublin Core specialties, DC: Collections. Although this had its attractions, it was believed that natural history collections would need certain specialised fields over and above what was under discussion for this developing standard – an example is the “Known to contain types” field.

The other development was the involvement of the NHM in the BioCASE project, funded by the European Union between 2001 and 2004 and led by Professor Walter Berendsohn of the Berlin Botanic Garden. The current BioCASE service can be found at http://www.biocase.org/

The BioCASE partnership of over 30 countries can be viewed as a precursor to GBIF, mobilising and integrating specimen data using new standards. It was recognised that it would be many years before all specimens would have a database record, so a means to alert researchers to collections with no database, but worthy of attention, was required. The data standard for the gathering and exchange of specimen data became Access to Biological Collections Data (**ABCD**) and the collection-level complement to ABCD became the foundation for NCD. Information

about ABCD can be found at
http://www.bgbm.org/tdwg/CODATA/Schema/default.htm and at
http://www.tdwg.org/standards/115/

National nodes were set up that would create and organise the data for their country and a central index was created from data harvested on a nightly basis from the national nodes. The Berlin team developed the National Node Data Input Tool (**NoDIT**), which was a Microsoft® Access™ database based on the collection-level data XML Schema, which enabled the system, using a computer-to-computer data transfer protocol that developed into TAPIR. A brief description of the harvesting process into the Core Metadatabase (**CorM**) can be found at
http://www.biocase.org/whats_biocase/meta_net_old.shtml

International interest in collections descriptions for natural history extended across the Atlantic with the formation of the RLG Natural History Group. **RLG** (which used to be known as the Research Libraries Group and is now RLG Programs, OCLC) acted as facilitator for several large natural history museums and botanic gardens, providing a mechanism for collaborative working that proved to be very successful under the leadership of Günter Waibel.

The Group became known as the RAVNS – Resources Available in Natural Sciences. This group developed the BioCASE collection-level XML schema from being relevant only to collections of specimens to a schema that would also manage descriptions of materials found in natural history libraries and archives, so making it truly cross-domain. This group was later expanded into the NCD Interest Group as part of the re-developed Taxonomic Databases Working Group (**TDWG**) and are collectively the authors of this standard.

Since TDWG mandated that the Resource Description Framework (**RDF**) would be its preferred technical environment, NCD was converted from an XML Schema into RDF. This proved to be a complex and controversial move and there are still calls for an XML Schema version, but keeping the two synchronised would be more troublesome than useful and so the normative version of NCD is in RDF, as presented here.

Appendix 2: Glossary

BCI	Biodiversity Collections Project. A central index of biodiversity collections around the world, based on NCD. http://www.biodiversitycollectionsindex.org/
BioCASE	Biological Collection Access Service for Europe. A multi-national specimen information network for Europe. http://www.biocase.org/
CorM	Core Metadatabase. Central database used for harvesting collection descriptions in the BioCASE project. http://www.biocase.org/whats_biocase/meta_net_old.shtml
CODENS	Institutional acronyms, abbreviations or other codes. See also http://circa.gbif.net/irc/DownLoad/kleYAJJ_moGCtjTxGtCbK1qGh-4pYxts/F-hH-dIxQfm2jIxFJFqGyi2s2wP/codenHowTo-v0.4.1-draft.html
DC	Dublin Core. Metadata element set that is a standard for cross-domain information resource discovery. http://dublincore.org/documents/1999/07/02/dces/
DCMI	Dublin Core Metadata Initiative. The organization engaged in developing Dublin Core metadata standard. http://dublincore.org/
EAD	Encoded Archival Description. The rich standard for encoding archival finding aids using XML. http://www.loc.gov/ead/
EDIT	European Distributed Institute of Taxonomy. Consortium to integrate taxonomic research. http://www.e-taxonomy.eu/
GBIF	Global Biodiversity Information Facility. Interoperable network of biodiversity databases and information technology tools. http://www.gbif.org/
JSON	JavaScript Object Notation. Lightweight data-interchange format. http://www.json.org/
NCD	Natural Collections Description is a data standard for describing collections. http://www.tdwg.org/activities/ncd/
NoDIT	National Node Data Input Tool. MS Access database used by the BioCASE National Nodes to record collection description data.
OCLC	Online Computer Library Center (previous name). A non-profit, membership, computer library service and research organization dedicated to the public purposes of furthering access to the world's information and reducing information costs. http://www.oclc.org/

PLANETS	Preservation and Long-term Access through Networked Services. Addressing core digital preservation issues. http://www.planets-project.eu/
PLATO	Preservation planning tool for digital objects, developed under the PLANETS project. http://olymp.ifs.tuwien.ac.at:8080/plato/website/intro.html
RAVNS	Resources Available in Natural Sciences. Group that was formed through RLG Programs to work on NCD.
RDF	Resource Description Framework. Lightweight ontology system to support knowledge exchange online. http://en.wikipedia.org/wiki/Resource_Description_Framework
RLG	RLG Programs. Formerly the Research Libraries Group, now part of OCLC. http://www.oclc.org/programs/about/default.htm
SYNTHESYS	Synthesis of Sytematics Resources. Large scale facilities EU Framework VI project. http://www.synthesys.info/
TDWG	Taxonomic Databases Working Group. Now known as the Biodiversity Information Standards (TDWG) group that develops standards and protocols for sharing biodiversity data. http://www.tdwg.org/
TGN	Thesaurus of Geographic Names. Developed by the Getty Institution, providing identifiers for place names. http://www.getty.edu/research/conducting_research/vocabularies/tgn/
URI	Unique Resource Identifier. Generic term for linking web resources, includes URLs. http://en.wikipedia.org/wiki/Uniform_Resource_Identifier
vCard	File format standard for electronic business cards. http://www.w3.org/TR/vcard-rdf
XML	Extensible Markup Language. A simple flexible text format playing an increasingly important role in the exchange of a wide variety of data on the Web. http://www.w3.org/XML/