DCA-Internal Science Team input for the Data Centre Census process

23 April, 2007

Action BOARD2-WP2-A3 asks the IST to comment on the process for gathering and presenting the information in the data centre census. This note provides recommendations from the IST.
We breifly outline the role of the census from the perspective of the IST. We discuss some over-all considerations, and we provide a set of recommended questions to be posed in the census.

Role of the Data Centre Census

We understand that the census has two main roles in the DCA project.

- Firstly as a major deliverable to EC in a form that clearly sets out the number, size, distribution and diversity etc. of data centres in Europe.
- Secondly, and more immediately relevant to IST, is that the census be used within the DCA project to capture basic information about existing data centres in order to decide on the topics for the DCA workshops, tutorials and data centre visits.

Overall Considerations

In commenting on the process for gathering and presenting the census information we have concentrated on the second role. In particular we consider what information should be included in the census in order for the DCA project to decide the best way to help the data centres in take-up of VO standards. And to help the DCA prioritize its effort in terms of potential scientific benefits.

Preliminary Census

The preliminary census provides a stating point, but is simplistic due to the limited information it contains. In particular the first two columns 'Data Archives/science ready data' and 'Added-value databases and services' are too general and mix up a number of different concepts. Critical information about data types, volumes and science readiness should be included. We try to address this with the recommendations below.

Also, the way the preliminary census is presented is not uniform. For example LEDAS (Leicester Database and Archive Service) which has some 3000+ catalogues plus several X-ray data archive, is presented as a single row entry, whereas each ESO instrument is presented on a separate row. Clearly there is hierarchy in the information which could better be presented by grouping rows, or by using some kind of tree structure.

Recommendations

1. Collect the information for the census directly from the data centres via a census form, with the ability to pose yes/no questions, multiple choice questions, questions with numerical answers. Also allow for limited textual answers and space for additional important descriptive information.

2. Separate the information about the archives from the information about the data/products

   For Archives the following information should be considered:

   - Total volume of holdings (GB)
   - Total number of astronomical objects, number of database rows and columns
     - For observation logs - number of pointings
   - Contact point and scientific contact point

- Uniqueness of the data - is the same data hosted elsewhere?
- Public/Proprietary nature
  - Completely open / Data public after proprietary period / Access restricted
- Human resources supporting the archive
- History and expected lifetime of the archive
- Is it a growing archive or are the contents static
- Number of users of the archive
- Level of VO awareness
  - current compliance with VO standards
  - intent to utilize VO standards
  - current interface - web form access to archive / ftp site / direct URL to data / other
  - is VO compliance required by the archive governing bodies?

For the data/products the following information should be considered:

- Type of data
  - Energy regime (Radio, mm, IR, Optical, UV, X-ray, Gamma-ray) [Here we suggest simply adopting same categories as defined for IVOA registries]

Radio	$\lambda \geq 10$ mm $\nu \leq 30$ GHz
Millimeter	0.1 mm $\leq \lambda \leq$ 10 mm
	3000 GHz $\geq \nu \geq$ 30 GHz
Infrared	1 $\mu \leq \lambda \leq$ 100 μ
Optical	0.3 $\mu \leq \lambda \leq$ 1 μ
	300 nm $\leq \lambda \leq$ 1000 nm
	3000 Å $\leq \lambda \leq$ 10000 Å
Ultraviolet	0.01 $\mu \leq \lambda \leq$ 0.3 μ
	100 Å $\leq \lambda \leq$ 3000 Å
	1.2 eV $\leq E \leq$ 120 eV
X-ray	0.1 Å $\leq \lambda \leq$ 100 Å
	0.12 keV $\leq E \leq$ 120 keV
Gamma-ray	$E \geq$ 120 keV

  - Classify data types as : Images / Spectra (single slit/ multiple slit/ IFU) / Interferometry / particle detections (event lists etc.) / other (specify)
  - Level of products [Using categories defined for IVOA registries]

A	Data are fully calibrated, fully documented, and suitable for professional research.
B	Data are calibrated and documented, but calibration quality is inconsistent. Users are advised to check data carefully and recalibrate.
C	Data are uncalibrated.
U	Data quality is unknown.

3. Collect more detailed information on Theory data centres. Distinguish between archives of theoretical models and services that provide computational capabilities.
   - Collect information on the general scientific area of the theory data centre. This could be classified using journal typ keyword, or simply the broad areas of the A&A keywords as listed below.
     - Physical data and processes

- Astrometry and celestial mechanics
- The Sun
- Solar system
- Stars
- Interstellar medium (ISM), nebulae
- The Galaxy
- Galaxies
- Cosmology

[Omitted items from A&A list: Sources as a function of wavelength; Astronomical instrumentation, methods and techniques; Astronomical data bases]

- Theory archives
  - For archives of pre-computed models collect the same information as for observational archive as above (where applicable).
- Theory services
  - What is the current interface: web interface / standalone tools / command line / other
  - Typical computation time (seconds/hours/days)
  - Does the service relate to simulations of observations ?
- Theory data products. [More input on possible types is needed here]
  - N-body cubes, spectra, images, other (specify)

4. Collect basic information about tools and software suites
   - Are the tools applicable to: images / spectra / particle detections / interferometry / catalogues / models / accessing other services / other
   - Name of tool
   - Contact point and scientific/technical contact point
   - What is the current tool interface: stand alone tools / web interface / command line / software environment / other
   - Level of VO awareness
     - Current compliance with VO standards
     - Intent to utilize VO standards
5. Collect basic information on Thematic Services

   Thematic services is a term being used in DCA in order to encompass almost any kind of service you might have in astronomy, making it difficult to pose specific questions here. As such we recommend collecting information on the general scientific area and intended interface to VO
   - Collect information on the general Scientific Area - As above for theory data centres
   - Specify whether it is a general purpose VO portal - i.e. a collection of services grouped by a theme or country
   - If possible specify the intended interface to VO
   - Describe the general type data products from or applicable to the service
     - Images / Spectra (single slit/ multiple slit/ IFU) / Interferometry / particle detections (event lists etc.) / other (specify)