Approaches to Taxonomy Development: Some Experiences in the Context of an Academic Institute Information Portal

Lohrii Kaini Mahemei, K Thulasi* and T.B. Rajashekar
National Centre for Science Information (NCSI)
Indian Institute of Science
Bangalore 560 012 (India)
{kaini, thulasi, raja}@ncsi.iisc.ernet.in

* Author to whom all correspondence should be sent

Abstract

SciGate, the science information portal of the Indian Institute of Science (IISc), Bangalore, organizes and provides access to a large number of online scholarly information resources to the Institute research community. Scope of content and information architecture of SciGate was defined and developed in 2001. The portal has consistent high usage. Formal and informal feedback from many users over the past three years of SciGate operation has necessitated us to revise the scope of SciGate content and its navigation interface. We needed an architectural approach for this redesign effort. 'Taxonomies' appear to offer a very useful approach to address the different design aspects of information portals. We formulated our approach based on a study of several operational portal taxonomies, their development and application. Our first two tasks were to define scope of our taxonomy based on user needs study and to develop the taxonomy outline. We describe details of our approach and results of the study. We also discuss our further plans towards implementation of the revised SciGate architecture.

1. Introduction

The Indian Institute of Science (IISc), Bangalore, is a premier institution of research and advanced instruction in India. It also has a very high international standing in the academic world. It provides facilities for post-graduate research and teaching in several areas of science and engineering. The Institute has more than forty academic departments, with over 2500 active researchers pursuing research, including about 500 faculty members. Each year IISc publishes close to 2000 research papers and awards about 200 research degrees. The Institute has excellent intranet and Internet connectivity. IISc researchers have online access to a large number of e-resources, including bibliographic and citation databases, data sets, over 10,000 e-journals and other web resources. Access to these resources is provided and managed through SciGate, the IISc science information portal and gateway web site (11). SciGate also provides access to several internal resources (e.g. IISc research publications) and open access online resources.

SciGate usage statistics indicate its consistent high usage by IISc users. However, recent formal and informal feedback from our users has required us to reconsider its architecture, both in terms of content and presentation. We needed an architectural approach for this redesign effort. 'Taxonomies' appear to offer a very useful approach to address the architectural needs of information portals. Based on a study of several operational portal taxonomies, their development and application, we have adapted taxonomy approach to redesign SciGate. We have found this approach quite useful. We are in the process of implementing the redesign. We describe our approach and results in this paper. In Section 2, we discuss the objectives of SciGate portal, its current design, and the need for redesign. In Section 3 we briefly review taxonomy-based approach to portal design. In Section 4 we discuss the application of this approach to redesign SciGate and findings. In Section 5 we discuss further work in implementing the new design.

2. SciGate information portal and need for redesign

The current architecture of SciGate information portal was designed during later half of 2001 and the portal was released for use in early 2002. The goal of SciGate is to act as a single point gateway to the IISc research community to a variety of locally hosted (internally produced or licensed) and Internet-based (free and licensed) science, engineering and management information resources.

SciGate Content: Core content of SciGate is meta-information of a large number of e-resources of relevance to IISc researchers. This metadata is presented to users in summary and detailed form through browse and search interfaces. Scope of e-resources covered in SciGate include: Commercial e-resources (e-journals, data sets, standards, etc.); open access scholarly e-resources available freely on the Internet of relevance to IISc researchers; IISc publications (research publications, multimedia, expert profiles, etc.); library catalogues and special information facilities in IISc; reference sources; and current science news.

SciGate user interface: SciGate provides a visually simple user interface for supporting browsing and searching its content. Browsing is supported through five catalogue tab-like pages (see Figure-1). Both simple and advanced searches are supported. Searches are carried out on the metadata stored in a backend MYSQL database.

[Figure-1]

SciGate backend: We use Dublin Core based metadata schema (see Table-1) for describing resources covered in SciGate. Metadata is managed using a MYSQL/PHP based content management system. This database backend is also used for supporting search, browse, and display. A link redirection program is used for capturing user clicks on e-resources (selected from browse and search results) and for gathering usage statistics.

Limitations of current architecture: Based on formal (gathered through SciGate 'feedback' form and analysis of usage statistics) and informal feedback from our users over the past three years of its operation, we have come to realize several limitations of the portal and improvements that need to be made. These are primarily related to SciGate content and its interface.

1. Scope of SciGate content. Focus of SciGate has mainly been scholarly e-resources of interest to higher education and research audience. Feedback from our users indicate their interest in accessing content from growing number of internal websites including administrative sections (e.g. academic, consultancy and research schemes units) and new facilities (e.g. IPR cell). Also, tendency of several IISc department websites to include links to external Internet resources of type not covered in SciGate (e.g. technical writing, effective use of software tools) seem to indicate the need for revising the scope of content in SciGate.
2. More 'direct' visualization of accessible resources. The current distribution of content by the five tab-like pages, though found useful, appears to be a source of some confusion among our users in resource identification, resulting in more mouse clicks. Though we have used metadata to describe e-resources, including subject classification and resource type category, this metadata has primarily been limited to support searching. There seems to be need for supporting a navigation mechanism that more tightly integrates access to the three main types of resources (free, licensed and internal) in the presentation interface.

In a knowledge-intensive environment like IISc, researchers require the best information possible for the kind of research they undertake and to identify these resources quickly and efficiently. Above mentioned considerations lead us to the need for redesigning the architecture of SciGate, in terms of scope of content covered, structure and navigation features. We find that the current design too much presentation-specific. It also does not make effective use of various categories in metadata (e.g. subject, resource type, availability) at presentation level. We needed an architectural approach for this redesign

effort. 'Taxonomies' appear to offer a very useful approach to address the architectural needs of information portals.

Table-1 SciGate Metadata

Field Name	Element Name and Name Space	Mandatory?	Definition	Source
Resource Title	dc:title	Y	Title of resource	
Author	dc:creator	N	Personal author	
Publisher	dc:publisher	N	Organization publishing/ hosting content	
Description	dc:description	Y	Summary of resource content	
URL	dc:identifier	Y	Online location	
Alternate URL	dc:identifier	N	Alternate location	
Subject Category	dc:subject	Y	Topic of content of the resource	SciGate subject categories
Resource Type	dc:type	Y	Types of resources covered	SciGate resource type categories
Availability	scigate:availability	Y	Access free or restricted to IISc	
Contact e-mail of publisher	scigate:contact	Y	-	
Resource Format	dc:format	N	Significant multimedia content	
Geographic Coverage	dc:coverage	N	If resource scope is limited to specific countries	
Time Coverage	dc:coverage	N	If resource scope is limited to specified time period	
Date record created	dc:date.created	Y	-	
Date record modified	dc:date.modified	Y	-	
Date link verified last	dc:date.linkverified	Y	Date resource last accessed	
Record created by	scigate:createdby	Y	-	

(staff)				
Record last modified by (staff)	scigate:modifiedby	Y	-	

3. Taxonomies and information portals

Historically used by biologists to classify plants or animals according to a set of natural relationships, there has been significant recent interest in using taxonomies for organizing content in information portals. We studied definitions and explanations given by several experts (2-4, 9, 13) and reviewed a few case studies (1-3, 5, 7, 10, 12) to assess relevance of taxonomy-based approach for our redesign effort. We found the following relevant constructs of taxonomies in the context of enterprise information portals.

- In essence, a taxonomy is a systematic way of organizing knowledge (e-resources, in our case). It provides a hierarchical structure of concept terms (categories and subcategories) that help in the development of a common language (vocabulary) to aid organization and sharing of knowledge. Typically, enterprise taxonomies are 'faceted' (e.g. communities, subjects, locations, etc.), each facet made up of hierarchical (or group of) concept terms. Facets thus allow taxonomies to be multi-dimensional.
- Taxonomies also provide a way to map alternate terms to root words that might include abbreviations, synonyms or other aliases.
- Enterprise taxonomies are usually customized to reflect the business processes, content assets, language, culture and goals of particular enterprise.
- Taxonomies and metadata work hand-in-hand. Metadata describes an asset in terms of a meaningful set of attributes. While much metadata is flat or one-dimensional in nature (e.g. title, author, creation date), some of it is taxonomy-based (e.g. subjects and content type).
- Taxonomies, combined with metadata, enable authoritative tagging of content during its creation and submission to the portal. The same framework is also used to provide support for browsing and searching. Taxonomy browsing enables users to understand both the scope and context of available information.

Taxonomies have thus been found to facilitate improved navigation and resource discovery, by using consistent metadata tagging, labeling schemes for navigation, browsing and searching (8). This is made possible by not only using most preferred terms as entry elements in navigation system, but also by using alternate term lists (controlled vocabularies) to manage synonym and homonym problems. Further, by using meaningful hierarchies, taxonomies also facilitate users to find resources at appropriate context.

Taxonomy Applications: Taxonomies, combined with metadata, have been applied in several information portals for effective content organization, search, browse and navigation. MSWeb, the Microsoft employee portal, is reported to use a taxonomy consisting of several vocabularies (e.g. subjects, product, geography, etc.), navigation labeling scheme and metadata schema (2, 10). This is used for tagging content at content creation stage, for site navigation (browsing), personalization and information retrieval (searching). NASA has evolved an agency-wide taxonomy framework for handling its electronic content and provides one stop shopping for NASA resources (3, 7). NASA information portal uses a faceted, hierarchical taxonomy (disciplines, functions, industries, locations, etc.), vocabulary control system, and Dublin Core based metadata scheme. Halliburton Company, one of the world's largest providers of products and services to oil and gas industries, developed a centralized taxonomy approach for consolidating multiple taxonomies that have been developed for separate product lines so that employees, contractors and customers can easily find information about products and services regardless of what product line they are in (1). Victoria Online is the Victorian government information portal. It is a metadata driven portal and uses Dublin Core based application profile and a taxonomy structure to support information and service discovery via searching and browsing (12). Victoria Online uses a set of taxonomies known as 'Facets' ('Topics', 'Do It Online' and 'Government Contacts'), each of which is a set of terms grouped and structured by a specialization/ generalization relationship.

Taxonomy Development: How are taxonomies developed? There is much useful experience available today for development and application of taxonomies for enterprise

information portals (3, 10, 13). Though taxonomy development methods vary in terms of details and specific techniques used, broadly following steps are involved:

1. Define scope of taxonomy
    - User need's survey: What content users need, how they access it
    - Information audit: existing content, its structure, who is responsible
    - Involve users: Include key stakeholders in the process
    - Identify existing vocabularies
2. Build and apply taxonomy
    - Identify needed attributes (facets), collect terms, build broad taxonomy outline
    - Review the taxonomy outline with stakeholders and subject matter experts and fill in the taxonomy outline
    - Develop tagging rules and procedures
    - Tag content
    - Expose content through the portal interface
3. Maintain taxonomy
    - Specify taxonomy maintenance business process
    - Document taxonomy maintenance procedures
    - Train users

4. Restructuring of SciGate: Our study and findings

In applying taxonomy approach to restructure SciGate, we focused on first two steps of taxonomy development: defining taxonomy scope and preparing taxonomy outline. In defining the taxonomy scope, we carried out following tasks to assess needs of our users:

- Study of select personal and department home pages
- Analysis of user profiles in the MySciGate personalized service
- Questionnaire-based survey of select faculty and research students

Through these studies, our goals were to:

- Identify and understand processes related to research, teaching and learning.
- Identify information requirements of our students, faculty and researchers in carrying out these processes.
- Identify and arrive at categories of information needs and resources

Study of select personal and department home pages:

IISc has over 35 departments and centres with faculty, researchers and students engaged in latest research in frontier areas. These departments fall under one of the six divisions (Biological Sciences, Chemical Sciences, etc.). Each department in IISc has its own website. Most departments include information regarding their staff, research areas, courses offered, research projects undertaken and so on. Many of these sites also include links to internal and external online resources. Based on this review we could identify several categories of information resources (Table-2).

Table-2 Category of information resources identified from department websites

Internal information resources	• Events -(Colloquium, Seminars, Conferences, Symposium, Workshop)
• People - Faculty, Research scholars, Students, Staff, Alumni	• Research Projects - (Sponsored, Consultancy)
• Courses (Regular, Short term)	• Administrative Principles and Policies (Rules and Regulations)
• Research areas	• Special Information Facilities/Dept Libraries.
• Facilities/Equipments/Instruments	***External Information resources***
• Employment opportunities at IISc	• Grant agencies
• Awards/Recognitions/Honours/Fellowships/Member/Endowments	• International collaborations
• Professional Bodies/Societies	• Institutions/Companies interacting with IISC
• Faculty's Research interests	• Web resources
• Research Publications - (Journal papers, Conference Papers, Thesis)	
• Admissions	

It may be seen from Table-2 that many of the resources are internal, and much of this content is not pure 'research' type. As of now, there is no centralized access point on the IISc intranet for finding and accessing these resources. Information related to online resources provided in the departmental websites is often incomplete or obsolete. Also, there is considerable duplication in online resources included among these websites. Expanding the scope of Scigate to provide up-to-date and well-organized access to such content seems to be of crucial importance to our users.

Analysis of user profiles in the MySciGate personalized service:

MySciGate is a web-based service designed specially for IISc researchers and is accessible from SciGate. Using MySciGate, users can create and maintain a customized profile web page, with links to their favourite electronic journals, databases and other web resources. MySciGate provides the researchers with a more efficient and effective means to monitor and access their favorite online information resources. There are around 200 profiles and analyzing these has helped us in understanding better the information requirements of our users.

From the analysis of MySciGate profiles, we identified following types of resources of interest to IISc researchers: E-journals, databases, newspapers, employment, search engines, software, institutions, facilities, conference sites, courses, e-print archives, mailing lists and research project sites. Since new profiles are regularly created and new resources added to existing profiles, MySciGate profiles would be a continuous source of design input to us.

***Questionnaire-based user survey*:**

We carried out a questionnaire-based survey among select faculty, researchers and students in IISc to get better insight into several aspects related to taxonomy design - processes they undertake; resources they require for handling these processes; resources they require for acquiring additional competencies and skills; and their information seeking behaviour in the present online environment. Separate questionnaires were

prepared for each community. The questionnaires were organized under the sections shown in the table-3.

Table-3: Survey Questionnaire Sections

Faculty Questionnaire	**Research Students Questionnaire**	**Masters Students Questionnaire**
1. Personal Information 2. Teaching and Research Processes 3. Teaching/Research tasks and information requirements 4. Resources for additional competencies/skills 5. Online resources accessed and usage 6. SciGate Categories & Tagging Schemes	2. Research processes 3. Research tasks and information requirements [1,4,5,6 as in column 1]	2. Learning processes 3. Learning tasks and information requirements [1,4,5,6 as in column 1]

The survey helped us to get better insight into specific processes IISc faculty, research and masters degree students undertake in fulfilling their core roles. In the context of taxonomy development for SciGate, the survey contributed significantly in delineating further the scope of content (e-resources) to be covered and also the categories under which these resources may be grouped. In Table-4 we provide a list of e-resources of interest to IISc researchers, based on the survey findings.

Table-4. Types of resources of interest to faculty and students

Information Resource Types	**Faculty**	**Research Student**
Resources for additional competencies	Scientific software, computer application software, resources on technical writing and	Scientific software, computer application software, statistical data analysis packages,

	communication skills, mathematical modeling, Presentation materials.	mathematical modeling, scientific datasets, sequence databases, structure databases, resources on technical writing and communication skills
Resources from other Institutions	Rarely	Yes, for example RRI, IIM, JNCASR, IIA, and IITs
Online resources access and usage	Online journals, bibliographic and citation databases, Institutional repositories, courseware, search engines, institutional websites	Online journals, bibliographic and citation databases, institutional repositories, search engines, institutional websites
Miscellaneous Resources	Railways, Airlines, Maps, Grant agencies, Government websites, Online bookstores	Grants agencies, Maps, Airways/Railways, Institutional websites, Online bookstores, Discussion forums, Tourist Guides
Information particularly needed but not accessible	Latest textbooks	Industry databases, Scientific databases, Company's technical documentation, Older issues of journals, Online Factual databases

Internal information to be included	Online directory/ phone book of all faculty, personnel, and academic and administrative units; all administrative policies and procedures (e.g. procuring equipments, attending conferences, duration for travel abroad); admissions; syllabus; scholarships/fellowship available; job opportunities; online facility to check status of bills and purchase orders; student information; academic information like research areas, research projects, and training courses; course materials; complete list of important events and dates (e.g. last date for sending in grade sheets, final exams, senate and faculty meetings); funding info; Estate office/complaint cell; Health center information; Research Schemes projects and statement of accounts; Information for new faculty e.g., forms for grant proposals, housing info, benefits/perks, start up grant info.

Based on the findings from the above tasks we have developed a draft taxonomy for use in SciGate. It is a faceted taxonomy, with 3 level hierarchy within each facet. This was obtained by bottom-up clustering of resource types based on attributes shared by these resource types. The table-5 shows the broad categories/facets of the draft taxonomy.

Table-5: Facets of the draft SciGate taxonomy

Facets	**Facet scope (illustrative)**
Organization	IISc Departments and Administrative units, major Indian research and academic institutions; grant sources, International/National Societies
Community	Staff including academic and administrative; Awards & honors bestowed; Alumni; Positions open in IISC
Activities	Courses offered, research areas, research projects, events in IISc
Information Sources	Internally generated academic and administrative information; Special Information Facilities; licensed and open access sources in Science, Engineering, Management

Services and Facilities	Internal Information services and facilities (academic and administrative, Personal)

Figure-2 shows the draft SciGate taxonomy.

[Figure-2]

5. Further work:

Much work remains to be done before we can launch the restructured SciGate portal. We need to take into account new content to be covered in SciGate identified during user needs survey. We need to extend the taxonomy with alternate terms for each of the root taxonomy terms. The taxonomy has to be embedded in the SciGate metadata scheme and tested by tagging sample content. More importantly we need to modify the SciGate navigation and search interface to exploit the taxonomy and metadata to integrate access to the three main types of resources (free, licensed and internal).

6. Conclusion:

In offering SciGate portal service, our goal has been to maximize the productivity of IISc faculty, researchers and students by enabling them to quickly identify available e-resources, internal or external, free or licensed. This should enable them to spend less time looking for useful information, spend more time in harnessing the discovered information to enhance their teaching, research and learning processes. We hope that the SciGate revision we are undertaking based on taxonomy-approach will help us in our efforts.

References

1. Busch, J.A. *Enterprise Information Architecture: A Framework for Intranet Success.* Retrieved November 10, 2004 from http://www.kmworld.com/kmw03/presentations/Busch.pps
2. Crandall, M. (2000). *Using Taxonomies Effectively in the Organization.* Retrieved October 20, 2003, from http://www.infotoday.com/KMWorld2000/presentations/crandall.ppt
3. Dutra, J. & Busch, J. (2003, January 7). *NASA Technical White Paper: Enabling Knowledge Discovery: Taxonomy Development for NASA*. Retrieved January 15, 2003, from NASA website: http://web-services.gov/NASA Taxonomy White Paper_final_rev.doc
4. Edols, L. (2001, October). Taxonomies are what? *Free Pint,* 97. Retrieved from http://www.freepint.com/issues/041001.htm#feature
5. Lider, B and Mosoiu, A. (April 2003). Building a Metadata-Based Website. *Boxes and Arrows.* http://www.boxesandarrows.com/archives/building_a_metadatabased_website.php
6. Mahemei, L.K. (January 2004). Approach to the development of a Taxonomy for SciGate (IISc Science Information Portal). *Major project report submitted in partial fulfillment for the Training Programme in Information and Knowledge Management,* National Centre for Science Information, Indian Institute of Science, Bangalore.
7. *NASA Taxonomy –Top Level Facets*. (2004). Retrieved June 10, 2004, from http://nasataxonomy.jpl.nasa.gov/
8. Rajashekar, T. B. (2003, August 9) *Taxonomies and Ontologies: An exploratory overview*. Retrieved September 17, 2003, from http://144.16.72.189/is206/taxonomies-tbr-srels.ppt
9. Ricci, C. (May 2004) Developing and Creatively Leveraging Hierarchical Metadata and Taxonomy. *Boxes and Arrows.* http://www.boxesandarrows.com/archives/developing_and_creatively_leveraging_hierarchical_metadata_and_taxonomy.php

10. Rosenfeld, L. & Morville, P. *MSWeb: An Enterprise Intranet #1*. Retrieved December 15, 2003, from http://www.boxesandarrows.com/archives/msweb_an_enterprise_intranet_1.php
11. *SciGate: The IISc Science Information Portal*. (2002). Retrieved August 15, 2003, from http://www.ncsi.iisc.ernet.in/
12. Victoria Online. (2002) *Victoria Online Metadata Application Profile (VOMAP).* Retrieved July 20, 2004, from http://www.egov.vic.gov.au/Victoria/VictoriaOnline/vomap.htm
13. Wyllie, J. *Taxonomies: Frameworks for corporate knowledge – the shape of things to come?* (2003). Ark Group.

Figure-1 Components of SciGate Interface

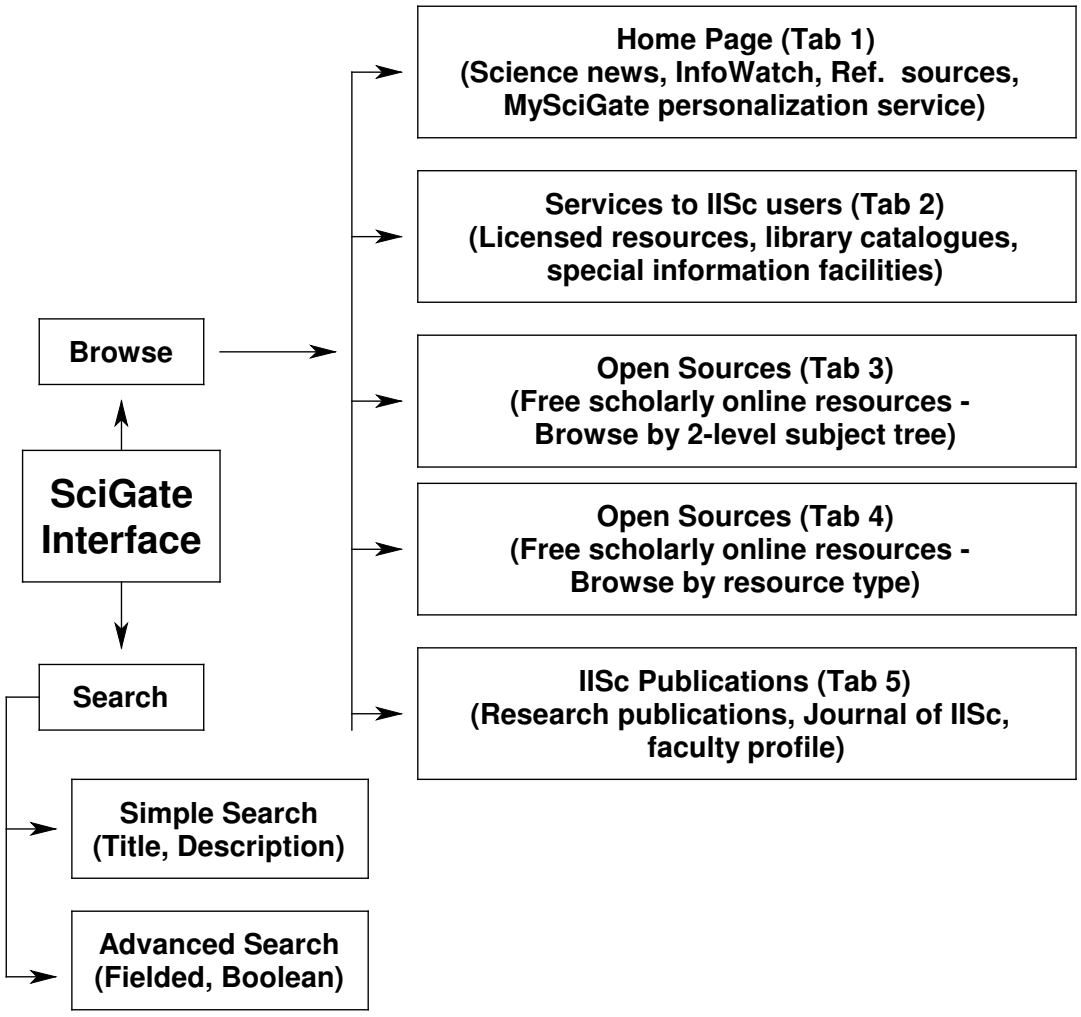
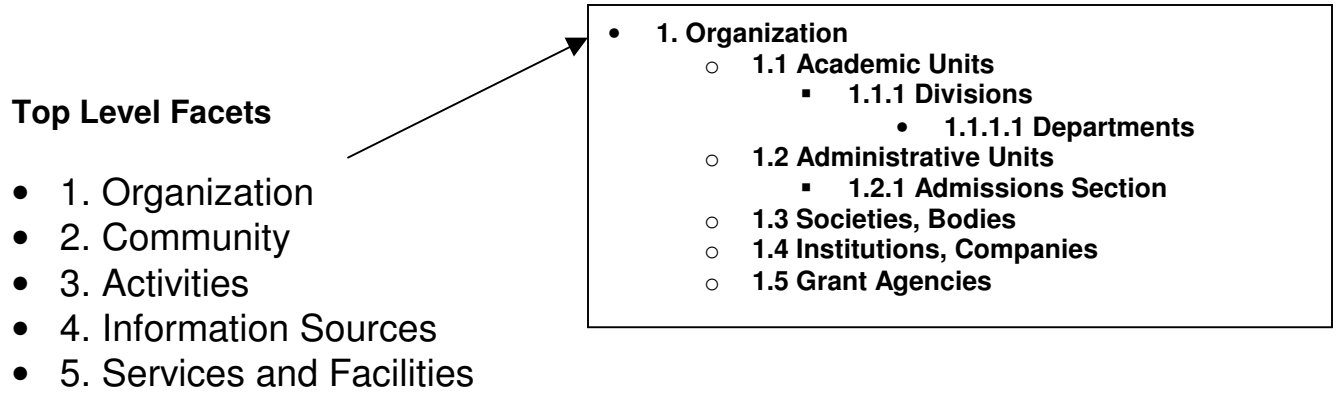


Figure-2. Draft SciGate Taxonomy



Top Level Facets (Expanded to 2 levels)

- **1. Organization**
  - 1.1 Academic units
  - 1.2 Administrative units
  - 1.3 Societies, bodies
  - 1.4 Institutions, companies
  - 1.5 Grant agencies
- **2. Community**
  - 2.1 People
  - 2.2 Awards and honours
  - 2.3 Positions open
- **3. Activities**
  - 3.1 Courses
  - 3.2 Major research areas
  - 3.3 Research projects
  - 3.4 Events
- **4. Information Sources**
  - 4.1 Licensed sources
  - 4.2 Internally produced sources
  - 4.3 Open access sources
  - 4.4 Special information facilities
- **5. Services and Facilities**
  - 5.1 Facilities and test centres
  - 5.2 Administrative services