Título de la Ponencia: Vocabularios controlados para la comunicación científica

Autores: Mgr. Diego Ferreyra. (Area Tecnología Documental CAICYT-CONICET.), Prof. Mela Bosch. Directora (CAICYT-CONICET)

Resumen

Los vocabularios controlados constituyen una herramienta para indispensable en el contexto de la comunicación científica. Aportan facilidades para la búsqueda, recuperación, visualización, análisis y representación comparada en cada campo y dominio científico. En tal sentido, disponer de una infraestructura de servicios semánticos en el CAICYT permite la comunidad científica argentina valerse de las herramientas necesarias tanto para optimizar y mejorar las condiciones de representación y descripción de recursos, como para expresar las líneas de investigación y desarrollo local, a la vez que establecer modelos de relación y análisis comparado con las agendas de producción científica a escala global. En la presente ponencia se describe el modelo de gestión establecido por el CAICYT para ofrecer una capa de servicios semánticos en el contexto de la comunicación científica en general y del aparato de ciencia y técnica Argentino en particular.

Palabras claves: COMUNICACIÓN CIENTIFICA, INFRAESTRUCTURAS DE INFORMACION, SERVICIOS SEMANTICOS, VOCABULARIOS CONTROLADOS

Indice:

Antecedentes y estado actual de los servicios de CAICYT
Infraestructuras de información orientadas a la comunicación científica
La Plataforma de Apoyo a la Comunicación Científica y Tecnológica Argentina
El Banco semántico
Conclusión
Bibliografía

Antecedentes y estado actual de los servicios de CAICYT

El CAICYT-CONICET tiene como cometido contribuir a la comprensión del desarrollo, evolución y transferencia de conocimiento mediante la investigación de la información publicada en ciencia y tecnología. Contribuye asimismo con servicios para la organización y acceso de la información científica y tecnológica y a la calidad de su difusión en las publicaciones nacionales

En los últimos años de uso intensivo de las tecnologías nos encontramos en un escenario que considera a la información como commodity, una mercancía destinada a uso comercial, como bien producido masivamente cuyo valor está en la relación oferta-demanda, sin diferenciación en función de quién o donde se produce.

Salir de esta presión es un desafío para los profesionales que formamos parte del ciclo de producción de información científica _ investigadores, autores, editores, bibliotecarios y documentalistas_ nos requiere convertirnos en desarrolladores de capacidades diferenciadas que no sean transformables en commodities.

Como institución esta meta nos lleva a mirar hacía el pasado y hacia adelante: el CAICYT tuvo su origen en 1958 como Biblioteca de CONICET, pasó a centro servicios documentales en 1976 fue promovido recientemente a unidad de investigación además de centro de servicios. Esto nos presenta un promisorio horizonte que nos impone enfatizar la consideración de la información como un producto social que se implementa en servicios a la comunidad altamente diferenciados.

Es por ello que nos preocupamos por individualizar quienes son nuestros usuarios, o mejor dicho los actores de la información científica argentina. Se trata de varios conjuntos: los especialistas de información, bibliotecarios, documentalistas, tanto del sistema de bibliotecas de CONICET como de otras instituciones y empresas interesadas en el desarrollo científico y tecnológico. Los editores de revistas científicas y en vinculación con ellos, los autores e investigadores que hacen de la información científica el objeto o insumo de su trabajo.

Considerando estos usuarios, al ritmo de las tendencias tecnológicas de las épocas, se han venido desarrollando servicios de tipo referenciales y de directorios para relevar la producción científica y tecnológica nacional, tales como el Catálogo Colectivo de Publicaciones Periódicas, la Guía de Unidades de Información en Ciencia y Tecnología. También en el aspecto de referencia y como identificador CAICYT es el Centro Nacional Argentino de ISSN desde 1974.

En los últimos años hemos además profundizado la línea de oferta del contenido sustantivo de acceso abierto: desde 2006 el CAICYT centraliza la participación argentina en SciELO (Scientific Electronic Library Online) red de virtual conformada por colecciones de revistas científicas en texto completo y con acceso abierto, libre y gratuito de países de Latinoamérica, España, Portugal y Sudáfrica. Las revistas que conforman la colección SciELO-Argentina son seleccionadas en cuanto a criterios editoriales y de contenido a través de un proceso de selección que se materializa en el Núcleo Básico de Publicaciones Científicas Argentinas. Es un proyecto del CONICET que identifica y distingue un conjunto de publicaciones científicas y tecnológicas argentinas en los distintos campos del conocimiento. Las revistas candidatas son sometidas a una evaluación exhaustiva con criterios de calidad y trascendencia. La evaluación de las revistas es realizada por un Comité designado por el Directorio CONICET compuesto por investigadores, tecnólogos, editores o docentes universitarios de reconocido prestigio, representando equitativamente las distintas áreas del conocimiento. Este Comité evalúa las revistas que son incorporadas al Núcleo Básico, y las reevalúa cada 3 años para certificar su permanencia en este proyecto. CAICYT es la sede

organizadora y de contacto para del proceso de selección. Un servicio complementario de CAICYT es el apoyo editorial con el Portal de Publicaciones Científicas y Tecnológicas, que sirve de semillero para iniciar la publicación digital.

Infraestructuras de información orientadas a la comunicación científica

La noción de infraestructuras digitales de información en el contexto de la comunicación científica hace referencia al conjunto de herramientas y medios digitales que sirven de soporte para las actividades de investigación y desarrollo en ciencia y tecnología.
La consolidación y uso de estándares, la disponibilidad de formas persistentes de identificación y representación de recursos en Internet y la generalización y convergencia de los medios digitales han afianzado las condiciones para la creación de nuevas prácticas y dinámicas en el campo de la comunicación social en general y de la comunicación científica en particular.

En los últimos años es posible observar el progresiva y creciente fortalecimiento de plataformas que exponen datos y servicios de manera persistente y uniforme, alcanzando niveles de previsibilidad y constancia que permiten que actores autónomos y variados desarrollen productos y servicios sobre los soportes funcionales y operacionales de las mencionadas plataformas.

De esta manera, a través de una lógica más performativa que normativa, muchas actividades científicas y tecnológicas utilizan como fundamento corriente servicios de datos públicos y privados provistos por terceros para el desarrollo de sus respectivas agendas de investigación y desarrollo.

En el campo de la comunicación científica son muchas las infraestructuras que intervienen y que han intervenido históricamente. En este sentido se ha avanzado la noción de infraestructura aplicada no sólo a estructuras físicas que dan soporte a otras actividades, sino también a dinámicas y modelos sociales que logran consolidarse como artefactos mediadores reconocidos socialmente.

Con este enfoque hacemos referencia a un conjunto de atributos propuestos por Bowker y Star (1999), para dar cuenta de una infraestructura en tanto artefacto mediador:

- Relación embebida: se encuentra inserta dentro de otras estructuras sociales, tecnológicas, etc.
- Transparencia: la infraestructura resulta transparente para el uso, a su vez, el soporte que brinda a otras tareas o estructuras resulta invisible.
- Ámbito de aplicación: tiene un alcance espacial o temporal que va más allá de un sólo evento o de una práctica aislada

- Aprendidas en el proceso de inclusión: la adopción y uso de una determinada infraestructura es una condición sine qua non para la pertenencia a una determinada comunidad de práctica.
- Que guarda relaciones entre convenciones de la práctica: la infraestructura tiene una naturaleza dual, constituye y es constituido por las convenciones de una comunidad de prácticas.
- Articulación con normativas: la infraestructura se articula de manera transparente a través de estándares y normas con infraestructuras y herramientas.
- Que está consolidada sobre una base previa: las infraestructuras no son creadas desde cero o como algo nuevo en su totalidad, siempre se articulan sobre bases de infraestructuras ya instaladas.
- Que se hace visible cuándo falla: el accionar normalmente invisible de la infraestructura se torna visible cuándo falla (ej: un apagón o la caída de un sistema).
- Que cambia de manera gradual: debido que la infraestructura resulta un estructura articular y compleja, provista de distintos significados locales, sus transformaciones no son globales ni tampoco se presentan como una cambio único y total. Sus transformaciones requieren tiempo, negociación y proceso de ajustes con respecto a los demás actores y sistemas involucrados.

Para explicitar más estos conceptos, y quizás haciendo una simplificación, nos permitimos reflexionar que así como la electricidad establece las condiciones de posibilidad de los electrodomésticos; aquellos servicios, prestaciones y capacidades cuya presencia, fiabilidad, utilidad y permanencia damos por cierta, son los que generan las condiciones de posibilidad para el desarrollo de nuevos servicios y prestaciones.

A este tipo de relación, de capacidad y transparencia la denominamos infraestructura. La infraestructura no sólo es un agente de legitimidad y autoridad, es un moderador de formas de participación e intervención social: cada infraestructura propone una arquitectura específica de participación, distribuye roles y posibilidades.

En contextos digitales, un concepto central para establecer y desarrollar infraestructuras de servicios esta dado por las capacidades de explotación y reuso. En gran medida estas capacidades están delimitadas por el esquema de portabilidad e interoperabilidad soportado por la infraestructura en tanto plataforma tecnológica de servicio.

En tal sentido, con el fin de adoptar una concepción amplia y comprehensiva de la noción de portabilidad, resulta útil la formalización conceptual realizada en el marco del Open Language Archives Community (OLAC). La misma considera que se encuentran cumplimentadas las condiciones de portabilidad cuando la información es codificada de manera tal que su uso, tratamiento y gestión pueda ser realizado:

1. utilizando diferentes aplicaciones informáticas;
2. utilizando diferentes infraestructuras informáticas;
3. según diferentes comunidades de prácticas;
4. en función de propósitos diversos;
5. a través del tiempo.

De esta manera, la propuesta de portabilidad del denominado marco OLAC está centrada en promover esquemas de codificación de datos que permitan usos no previstos, por parte de usuarios no previstos y en contextos no previstos en un primer momento por quiénes han elaborado los datos. Justamente, uno de los atributos más valiosos de las obras de infraestructura como tal es su capacidad para tolerar y viabilizar un amplio abanico de estrategias de uso y explotación, a veces incluso no previstas en el esquema inicial del servicio.

La Plataforma de Apoyo a la Comunicación Científica y Tecnológica Argentina

A través de un concepción amplia de la comunicación científica en tanto forma singular y especializada de una relación dinámica entre condiciones de producción, circulación y recepción de discursos, y con el fin de contribuir a la conformación de una infraestructura de servicios de información útiles a las actividades de investigación en ciencia y tecnología, el CAICYT-CONICET se propone establecer una **Plataforma de Apoyo a la Comunicación Científica y Tecnológica Argentina** a través de la cual se ofrecerá a investigadores, editores, bibliotecarios y comunicadores científicos un acceso al conjunto de servicios desarrollados en condiciones técnicas que permitan maximizar su uso, explotación y re-uso.

La **Plataforma de Apoyo a la Comunicación Científica y Tecnológica Argentina** se cnecuentra centrada en 4 ejes de trabajo:

1. **Apoyo Competencias para la gestión de la Comunicación Científica y Tecnológica:** Implica la puesta en disponibilidad de servicios operacionales como ayuda, por ejemplo, al manejo de citaciones, así como tutoriales y materiales de autoformación que pueden utilizar los investigadores.

2. **Consolidación de las infraestructuras de información**: Esto implica la disponibilidad uniforme y persistente de estructuras de descripción estables (metadatos) e instrumentos de denominación apropiados. (listas de valores, vocabularios controlados, tesauros, ontologías) lo cual permitirá expresar las líneas de investigación y acción científica del CONICET en forma consistente y favorecer el diálogo entre agendas científicas a nivel global.

3. **Promoción de entornos de exposición de la Comunicación Científica y Tecnológica**. Se trata de proveer de medios técnicos acordes a los actuales modelos mostración y circulación de publicaciones en la comunidad científica global. Se suministran los medios técnicos de asignación de metadatos para digitalizar recursos así como proveer a su persistencia en repositorios y en el desarrollo de flujos de documentales no redundantes, significativos y reusables.

4. **Apoyar a la creación de instrumentos de autoconocimiento del sistema científico**. Propende el desarrollo de métricas útiles para la caracterización de la producción formalizada en ciencia y técnica. Esta línea de trabajo está orientada a proveer de instrumentos para la toma de decisiones como así también disponer de modelos que permitan condiciones de comparabilidad entre actores científicos y con respecto a la propia actividad de CONICET en el curso del tiempo.

Continuando con las paralelos de ejemplificación, al igual que las autopistas o los trenes, para alcanzar un modelo de explotación parangonable a la noción de infraestructura no alcanza solamente con ofrecer amplias capacidades o prestaciones críticas, los datos deben ser provistos de manera persistente, uniforme y homogénea, es decir, deben ofrecerse en condiciones de previsibilidad constante para que otros actores sociales puedan diseñar y articular servicios basados en la infraestructura, en esta caso, provista por el Estado.

Volviendo al tema que proponemos, el uso de vocabularios controlados, en tanto conjunto de formalizaciones lingüísticas orientadas a controlar y delimitar el alcance de un término con el fin de proveerlo de condiciones para la identificación, diferenciación y comparación entre denominaciones, conceptos, eventos y/o entidades sociales, resulta una herramienta esencial y crítica en el contexto de las actividades de investigación y desarrollo científico y tecnológico.

Las herramientas de control terminológico, son artefactos mediadores que dinamizan y permiten los procesos de comunicación entre dominios, proveen un soporte que permite construir edificios y puentes conceptuales utilizando términos y relaciones lógicas.

Esta tarea, tal como señalan Bowker y Star (1999), es realizada por los vocabularios controlados a través de tres grandes tareas: establecer condiciones de comparabilidad semántica entre entidades, definir lo visible y lo invisible (aquello que no puede ser categorizado es invisible) y controlar y definir el alcance de un término.

La articulación de estas capacidades permite disponer y ofrecer un conjunto de servicios semánticos relevantes tanto para la búsqueda y representación de documentos, como así también para la identificación de patrones, extracción de palabras claves, reconocimiento de entidades y/o conceptos, clasificación automática de textos, y diversos esquemas descubrimiento aplicables a grandes bancos de recursos.

Adicionalmente al de disponer servicios semánticos capaz de proveer términos y definiciones de manera interoperable es posible su articulación con otras plataformas tecnológicas, como ser repositorios digitales o ampliaciones utilizadas para la gestión de la actividad científica y tecnológica ofreciendo vías de complementación entre procesos de gestión y recursos. Estas capacidades revisten utilidad directa para bibliotecarios e investigadores como así también para los demás actores que participan en la comunicación científica.

El Banco semántico

Banco Semántico que está desarrollando CAICYT en el marco de la Plataforma de Apoyo a la Comunicación Científica está orientado a desarrollar y proveer servicios identificación y representación de conceptos a través de términos y relaciones lógicas.

Para su implementación en tanto plataforma informática de servicios, se han definido un conjunto inicial de requerimientos generales técnicos y funcionales. Los presentamos a continuación separados según sean requerimientos de gestión o de servicios y explotación.
Características técnicas y funcionales para la gestión:

- Permitir la gestión de un número ilimitado de vocabularios controlados.
- Permitir la gestión distribuía de los vocabularios controlados
- Permitir esquemas de gestión multi-usuario con roles diferenciados de gestión.
- Permitir la implementación de diversas políticas y dinámicas de gestión (centralizada, distribuida, federada, etc.).
- Disponer de herramientas y rutinas de auditoría y control de calidad terminológico.

- Disponer de mecanismos activos de control de consistencia e integralidad de datos terminológicos (términos duplicados, control activos de relaciones semánticas no permitidas, control activo de concurrencias no admitidas, *workflow* de gestión, etc.).
- Disponer de funcionalidades orientadas a facilitar el trabajo distribuido (sugerencia de términos, trazabilidad, servicios de alerta, comunicación interna, etc.).
- Permitir procesos de importación masiva de términos y vocabularios controlados.
- Permitir definir metadatos descriptivos globales y autónomos para cada vocabularios controlado.
- Permitir la gestión independiente de usuarios según cada vocabulario controlado.
- Contempla vías y modelos de exportación en los principales esquemas de metadatos en uso (Skos-Core, Zthes, TopicMap, Dublin Core)
- Permitir realizar impresiones completas de cada vocabulario controlado.
- Permitir la creación de tipos de relaciones terminológicos definidos por el administrador de cada vocabulario controlados.
- Permitir la creación de tipos de notas definidas por el administrador de cada vocabulario controlados.
- Permitir definir tipos de relaciones entre el vocabulario y otras fuentes estructuradas de datos (linked data).
- Permitir la creación y gestión de vocabularios multilingües.
- Disponer de reportes de auditoría y control de calidad

Con respecto a la gestión terminológica en particular:

- Permitir la creación ilimitada y edición de términos, relaciones terminológicas y notas.
- Disponer de herramientas para la detección y auditoría de errores (términos duplicados, polijerarquías, términos libres, términos sin relaciones jerárquicas)
- Disponer de herramientas activas de control lógico (cumplimiento de reglas lógicas definidas para el vocabulario controlado en tanto ontología terminológica)
- Disponer de herramientas para el mapeo terminológico entre vocabularios controlados.
- Contemplar la posibilidad de definir relaciones entre términos de diferentes vocabularios controlados
- Contemplar la posibilidad de definir relaciones entre términos de un vocabulario controlado y una entidad web

Características técnicas y funcionales para la explotación de servicios

- Ofrecer un modelo de búsqueda básica y un esquema de búsqueda avanzada que contemple todas la entidades de gestión de cada vocabulario controlado.
- Disponer de mecanismos de expansión de búsqueda
- Disponer de mecanismos para implementar servicios de texto predictivo
- Disponer de mecanismos para la recuperación de búsquedas con resultados nulos.
- Ofrece la posibilidad de acceder a representaciones de metadatos terminológicos en los diversos formatos
- Dispone de una interfaz de servicios web terminológicos
- Deberá ofrecer representaciones de metadatos terminológicos compatibles con los criterios de buenas prácticas de la iniciativa Linked Data
- Deberá disponer de un punto de consulta SPARQL.

En virtud de los requerimientos antes mencionados, se adoptó la herramienta web TemaTres para la implementación informática del ambiente de gestión del banco de semántico. La interfaz pública de consulta y explotación de datos terminológicos se desarrolló utilizando la interfaz de servicios web provista por la herramienta.

Para el Banco Semántico el CAICYT propone una forma de localización basada en la infraestructura informática y en la convocatorio a referentes científicos de cada vocabulario, que pueden ser científicos del ámbito de CONICET o bibliotecas, editores o centros de investigación interesados en el uso del recurso.

El estado actual del trabajo contempla:
Ya realizado: Implementación servidor de vocabularios como la plataforma del Banco Semántico según las mayor parte de las funcionalidades y requerimientos arriba presentados.

En el aspecto gestionario la agenda próxima se encuentra focalizada en:

1. Seleccionar, adaptar, adoptar y localizar vocabularios controlados para cada dominio en el contexto de la agenda científica del CONICET, lo que requiere:
    - Definir macro-dominios y dominios prioritarios
    - Seleccionar interlocutores terminológicos
    - Conformar comunidades de gestión según dominio
    - Relevar estructuras de representación del conocimiento en uso
    - Extender y mapear vocabularios

- Definir esquemas de reuso y explotación

2. Analizar los de dominios temáticos a cubrir. Ya están propuestos como dominios iniciales:
   - Ciencias de la información
   - Matemáticas
   - Astronomía y Astrofísica
   - Física
   - Química
   - Ciencias de la Vida
   - Ciencias de la Tierra
   - Ciencias Agrarias
   - Ciencias Médicas
   - Ciencias Tecnológicas
   - Antropología
   - Ciencias Económicas
   - Historia
   - Ciencias Jurídicas y Derecho
   - Lingüística
   - Educación
   - Ciencia Política
   - Psicología
   - Sociología
   - Filosofía

Conclusión

La actual línea institucional de CAICYT convoca al protagonismo de los actores de la comunicación científica, investigadores, autores, editores, basado en la colaboración para la revalorización a través de la tangibilización de los recursos en información científica en Argentina.

Nuestra estrategia se basa en incidir en la cadena de valor con mayor calidad. Ponemos el énfasis en los vocabularios controlados como recursos semánticos claves en los procesos de búsqueda y representación a la vez que para viabilizar esquemas descubrimiento de nuevo conocimiento. Esto

implica hacer disponibles infraestructuras de información capaces de dar espacio a la construcción colaborativa.

Con esta visión estamos decididos a hacer evidentes los atributos y beneficios de nuestros prestaciones, a mostrarnos como referente estable y consistente, a ser reconocidos por nuestro nombre, a extender nuestros servicios con bienes complementarios y finalmente a ser sustento de una infraestructura de recursos apoyada en personas, profesionales abiertos a la comunidad con calidad y responsabilidad social.

Bibliografía

Audilio Gonzales-Aguilar, María Ramírez-Posada y Diego Ferreyra (2012) TemaTres: software para gestionar tesauros. En: El profesional de la información, 2012, mayo-junio, v. 21, n. 3.

Bird, S. & Simons G. (2003) Seven dimensions of portability for language documentation and description. Language, 79(3), 557-582. Fecha de consulta 2013-08-01, de: http://www.language-archives.org/documents/portability.pdf

Bowker, G. C., Star, S. L. (1999). Sorting things out Classification and its consequences. Cambridge, Mass: MIT Press.

Bowker, Geoffrey C., et al. (2010). Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment. En: J. Hunsinger et al. (eds.) International Handbook of Internet Research. Springer Science+Business Media.

Ferreyra, Diego (2013). ¿Cómo desarrollar bienes y servicios públicos con datos? En: Gestión municipal y gobierno electrónico: participación, transparencia y datos abiertos. Buenos aires: BID

IFLA (2013) Riding the Waves or Caught in the Tide? Navigating the Evolving Information Environment. Disponible en: http://trends.ifla.org/files/trends/assets/insights-document_cover_large.png Fecha de consulta 2013-10-10

ISO 25964-1:2011. Information and documentation: Thesauri and interoperability with other vocabularies (Part 1: Thesauri for information retrieval)

National Information Standards Organization (U.S.). (2005). Guidelines for the construction, format, and management of monolingual controlled vocabulary. National information standards series. Bethesda, Md: NISO Press.

O'Reilly, T. (2005) Qué es la web 2.0: Patrones del diseño y modelos del negocio para la siguiente generación del software. Fecha de consulta 2007-08-26, de : http://sociedaddelainformacion.telefonica.es/jsp/articulos/detalle.jsp?elem=2146

Principles of Open Government Data (2007) Disponible en:
https://public.resource.org/8_principles.html Fecha de consulta 2012-10-03
Swan, Alma (2012) Policy Guidelines for development and promotion of Open Access. UNESCO: París
Veron, E. (2004). Fragmentos de un tejido. Barcelona: Gedisa
W3C. Oficina española. Guía Breve de Servicios Web. Disponible en:
http://www.w3c.es/Divulgacion/GuiasBreves/ServiciosWeb Fecha de consulta 2012-10-03
Wright, Alex (2007) Glut: Mastering Information Through The Ages. Joseph Henry Press, 2007