INTEGRACIÓN DE REPOSITORIOS SEMÁNTICOS

Un camino hacia los datos abiertos enlazados

Renato Mazzanti, Carlos Buckle, Marcos Zárate, Gustavo Samec

INTRODUCCIÓN

- Ley 26.899 de Repositorios Digitales Institucionales de Acceso Abierto. 2013, 2016.
- Sistema Nacional de Repositorios Digitales (SNRD) 2013. Sistema Nacional de Datos Biológicos (SNDB) 2010 y el Sistema Nacional de Datos del Mar (SNDM) 2012.
- Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB)
- El proyecto: "*Infraestructura de Acceso a Datos Primarios con aporte de semántica en Repositorios Digitales*"
- El desafío significa resolver aspectos de interoperabilidad a nivel semántico y uniformidad de vocabularios, para lo cual es necesario repensar los modelos clásicos de organización del conocimiento en RDIs.

WEB SEMÁNTICA

- La Web Semántica es una extensión de la web actual en la que la información contiene un significado bien definido, lo que permite a las computadoras y las personas trabajar en cooperación.
- Berners-Lee, Handler, Lasilla:The Semantic WebIn: Scientific American 284.5, 2001.
- Una Web Semántica es una red de datos que pueden ser procesados directa o indirectamente por las máquinas.

DATOS ENLAZADOS

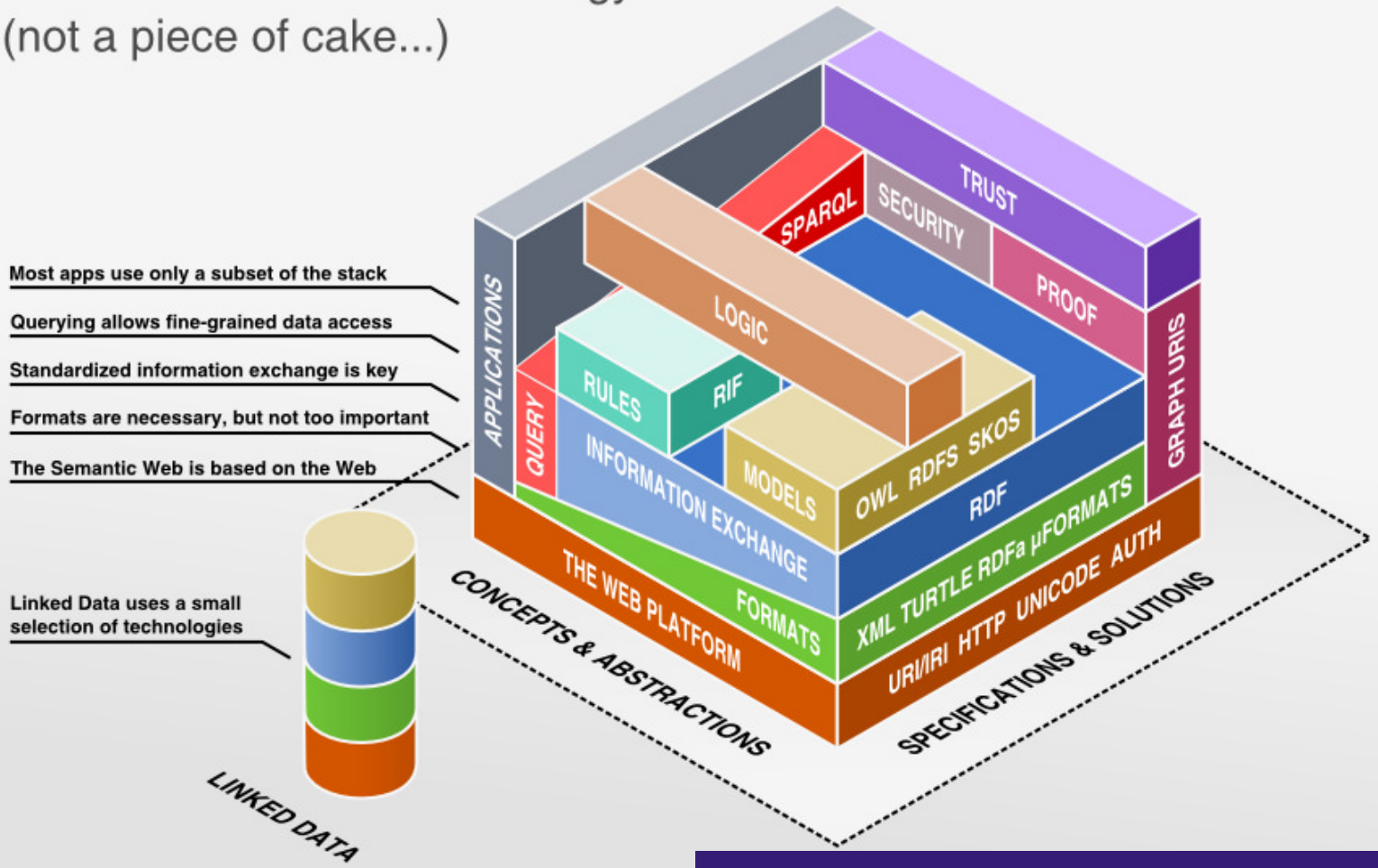
- El objetivo de la iniciativa de ***Datos Enlazados*** es crear una ***Web de datos*** con ***relaciones explícitas y semánticas entre los mismos.***
- La idea del ***razonamiento automático*** es parte de la idea original de la Web Semántica.
- Utiliza el estándar ***Marco de Descripción de Recursos*** (Resource Description Framework, RDF) para representar tanto los datos como las conexiones entre ellos.

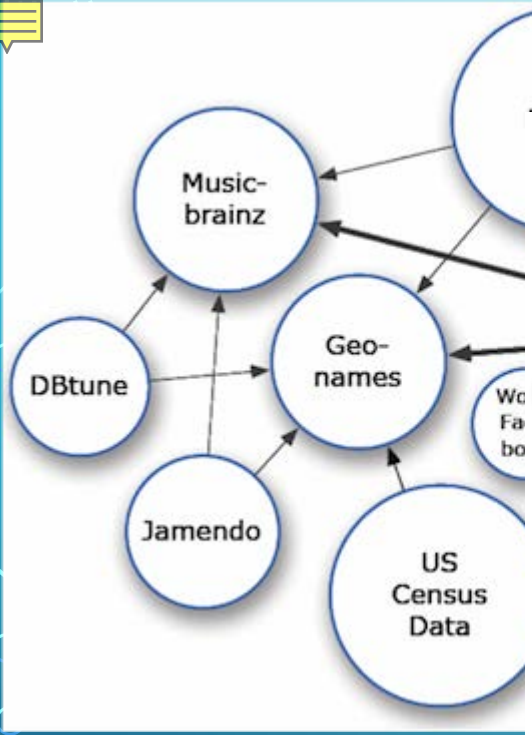
LOS PRINCIPIOS DE LOS DATOS ENLAZADOS

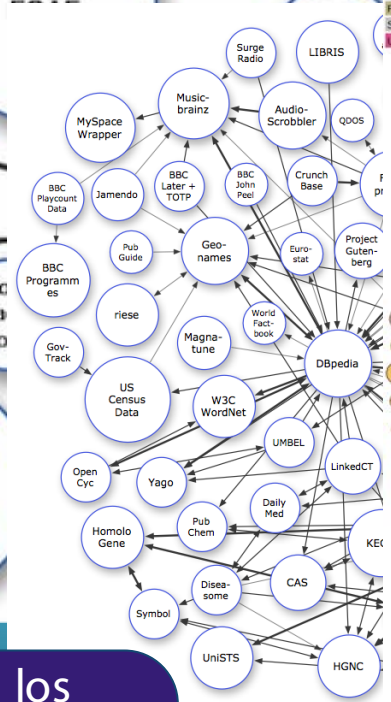
1. Utilizar URIs (Uniform Resource Identifier) para nombrar las cosas (recursos).
2. Utilizar URIs con HTTP para que las personas puedan buscar esas cosas.
3. Cuando alguien obtiene una URI, utilizando los estándares (RDF *, SPARQL), esta proporciona información útil.
4. Incluye enlaces a otras URI, para que se puedan descubrir más cosas.

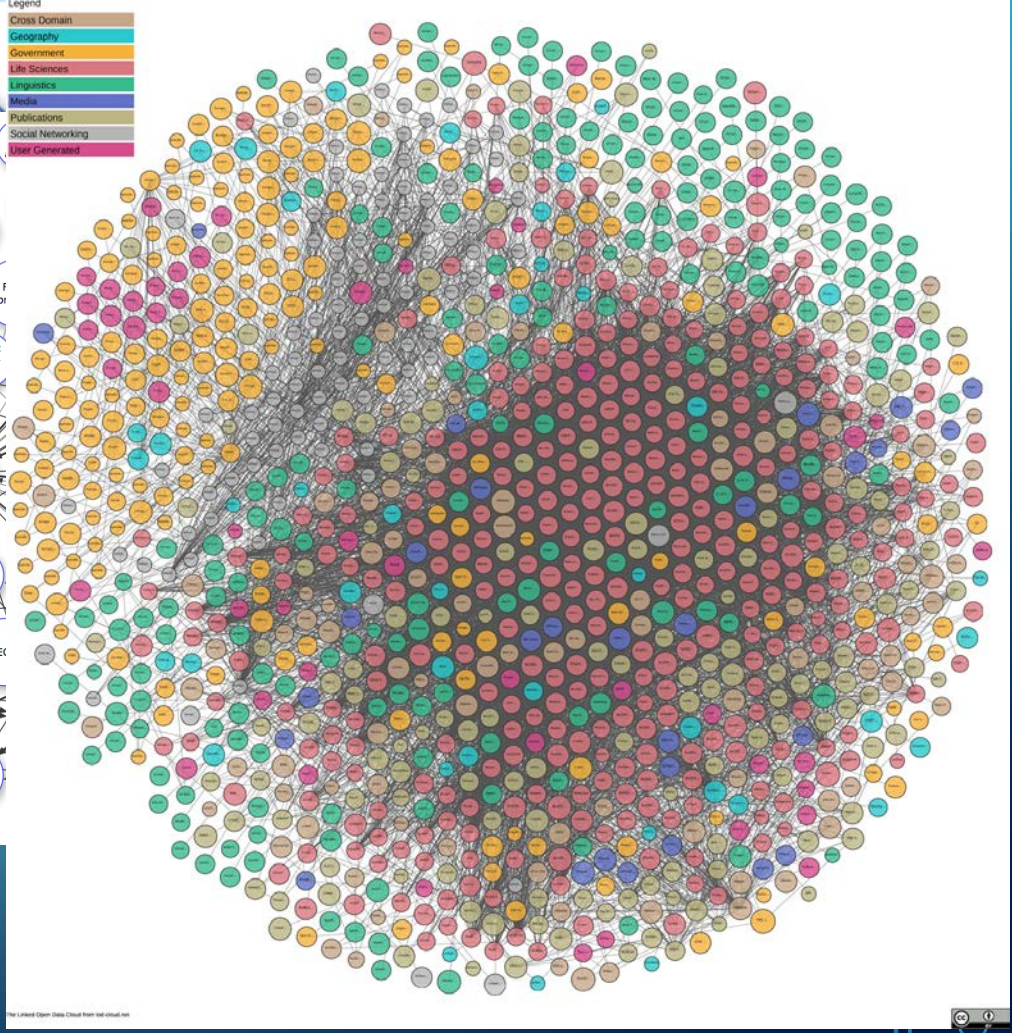
Tim Berners-Lee http: // www.w3.org/DesignIssues/LinkedData.html

The Semantic Web Technology Stack
(not a piece of cake...)









Esta imagen muestra los conjuntos de datos que se han publicado LOD.

https://lod-cloud.net/

¡Los contenidos de los repositorio son particularmente adecuados!

- Los metadatos ya existen en forma estructurada.
- No tienen que ser generados o ingresados manualmente para publicación como Datos Enlazados

- Los contenidos del repositorio se verán como datos procesables

Pascal-Nicolas Becker:
https://wiki.duraspace.org/display/~pbecker/Bringing+DSpace+into+the+Semantic+Web

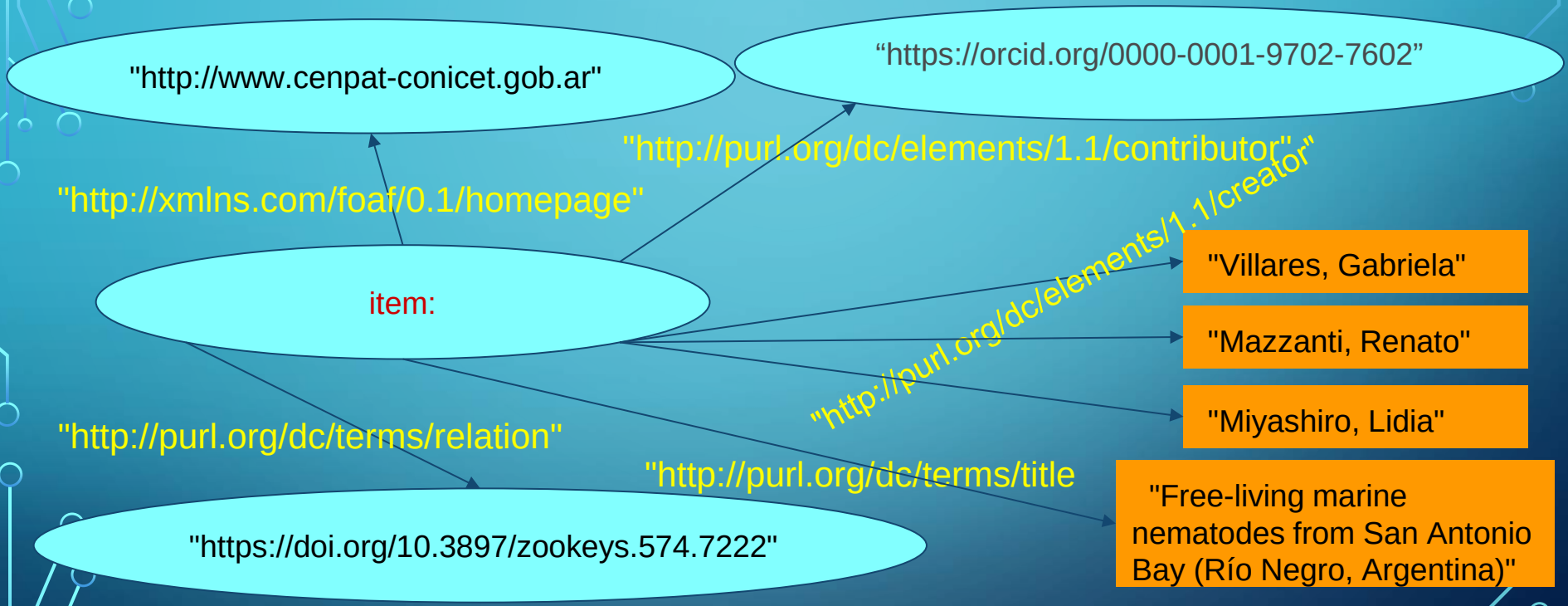
¿Qué pasa con OAI-PMH?

- Open Archive Initiative – Protocol for Metadata Harvesting es el estándar de hecho en el contexto de los RI.
- Google retiró el soporte para OAI-PMH en 2008.
- “Solo” es una interfaz limitada al contexto de los repositorios.

Limitaciones de los datos enlazados

- DSpace puede cosechar otros repositorios usando OAI-PMH.
- El soporte de datos enlazados en DSpace actualmente está orientado solamente a la exportación.
- OAI-PMH puede cosechar todos los documentos modificados en un intervalo de tiempo especificado. En Datos Enlazados, todavía hay que acordar vocabularios y/o convenciones.

Resource Description Framework, RDF

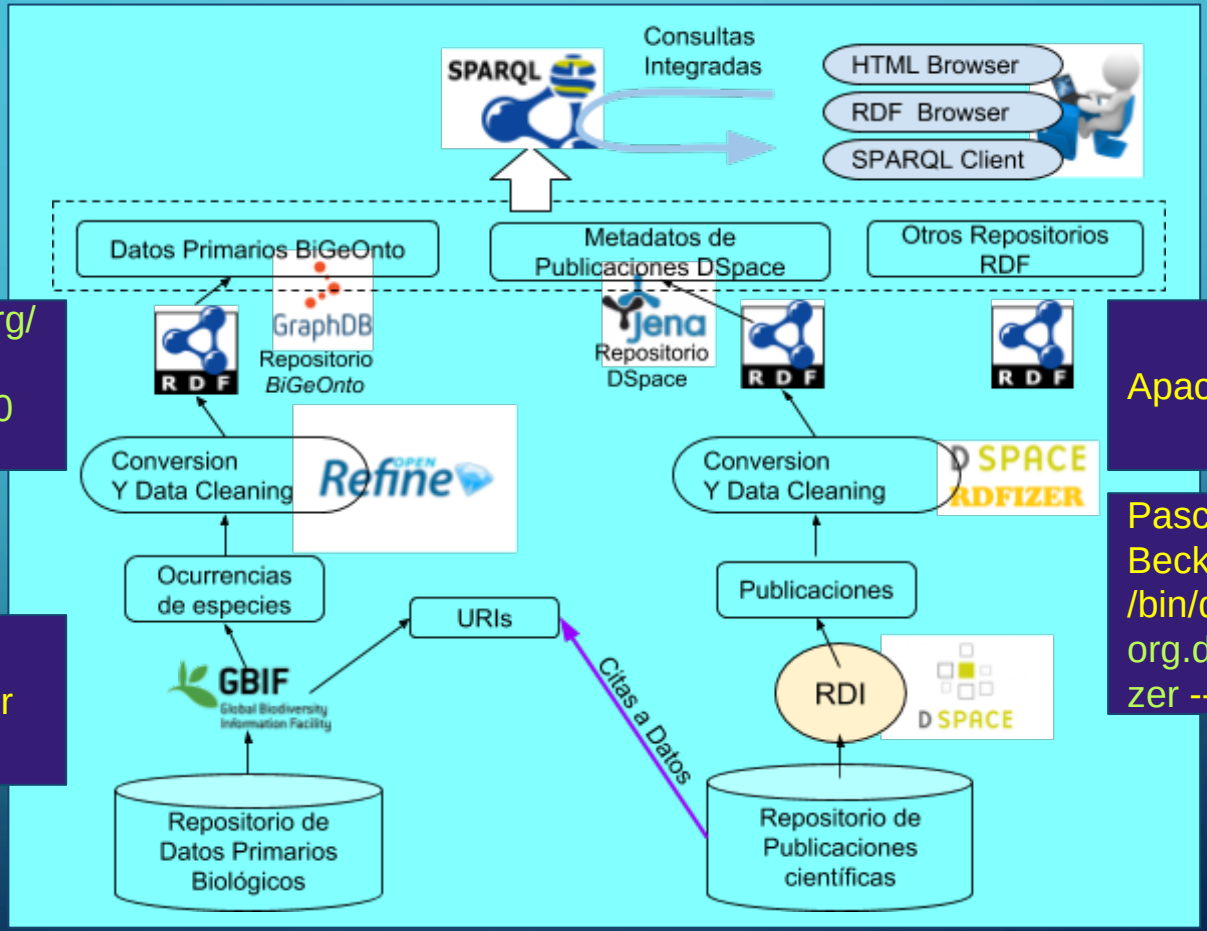


item:"http://www.repositorio.cenpat-conicet.gob.ar/rdf/resource/123456789/938"

¿Por qué datos enlazados en lugar de OAI-PMH?

- Los datos enlazados son una forma genérica y nativa de intercambio de datos.
- No limitado al campo de los repositorios.
- Los datos publicados siguiendo los Principios de Datos Enlazados son autodescriptivos.
- Los datos enlazados simplifican el intercambio de datos, con otros repositorios, para todos los datos que están fuera del entorno del repositorio al que pertenecen.
- Los “SPARQL (Protocol and RDF Query Language) end points” permiten búsquedas dentro del contenido de repositorios externos.

Arquitectura e implementación



http://www.w3id.org/cenpat-gilia/bigeonto/1.0.0 (OntoUML)

Internationalized Resourse Identifier (IRI) dwciri:

Apache Jena Fuseki

Pascal-Nicolas Becker /bin/dspace dsrun org.dspace.rdf.RDFizer --help

Consulta de Integración

#Recuperar las publicaciones de los RDI que han utilizado #datosprimarios relacionados a Puerto Madryn, Chubut".

```
PREFIX dc:
<http://purl.org/dc/terms/>
PREFIX bigeonto:
<http://www.w3id.org/cenpat-gilia/bigeonto/>
PREFIX dwc:
<http://rs.tdwg.org/dwc/terms/>
PREFIX geo-pos:
<http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX geo:
<http://www.opengis.net/ont/geosparql#>
PREFIX
dcelem:<http://purl.org/dc/elements/1.1/>
```

#Esta parte de la consulta recupera los DOIs de los conjuntos de
#datos primarios que tienen ocurrencias en Puerto Madryn.

```
SELECT ?titledspace ?publisher ?date
WHERE {
?s a dwc:Occurrence.
?s bigeonto:memberOf ?dataset.
?dataset dc:identifier ?identi.
?s bigeonto:has_event ?event.
?event bigeonto:has_location ?location.
?location dwc:locality ?locality.
# Filtramos por localidad
FILTER(regex(str(?locality),"Puerto Madryn" )).
# Filtramos solo los DOIs
FILTER(
strstarts(str(?identi),"https://doi.org")).
```

Invocamos al endopoint jena (respositorio DSpace)

```
SERVICE
<http://200.41.229.219:3030/dspace/sparql> {
        ?item dc:title ?titledspace.
        ?item dcelem:publisher ?publisher.
        ?item dc:issued ?date.
        ?item dc:relation ?url.
# Filtramos los DOIs de las publicaciones
FILTER ( strstarts(str(?url), "https://doi.org") ).
# Comparamos si existen DOIs iguales en ambos repositorios
FILTER (?url = ?identi )
}
}
GROUP BY ?titledspace ?publisher ?date
LIMIT 10
```

http://web.cenpat-conicet.gob.ar:7200/sparql?savedQueryName=datos-primarios

Resultado de la consulta

titledspace	publisher	date
El pingüino de Magallanes. I.- Evaluación y estratificación de densidades de su población en Punta Tombo, Chubut, Argentina	Consejo Nacional de Investigaciones Científicas y Técnicas - CONICET, Secretaria de Estado de Ciencia y Tecnología - SECYT, Centro Nacional Patagónico CENPAT	1979
El pingüino de Magallanes. I.- Evaluación y estratificación de densidades de su población en Punta Tombo, Chubut, Argentina	Centro Nacional Patagónico CENPAT - Consejo Nacional de Investigaciones Científicas y Técnicas CONICET	1983
Dinámica poblacional del elefante marino del sur (Mirounga leonina) en Península Valdés, Chubut	Universidad Nacional de La Plata	1988-02-28
Demersal and pelagic species of fish and squid from the Patagonian shelf	Pensoft Publishers	2017-04-13

CITAS A DATOS ABIERTOS (OPEN DATA CITATION)

- La forma más habitual de ***referenciar los datos primarios*** en una publicación es a través de una declaración de acceso a datos.
- **No se cumple con varios aspectos:**
  - si hay un error tipográfico en el identificador o URL, no hay información adicional para ubicar los datos entre las existencias del repositorio;
  - los autores pueden tener la tentación de dar la URL del repositorio, en lugar de una cita específica para el conjunto de datos;
  - no da el debido crédito a los creadores del conjunto de datos, un punto especialmente importante si estos son diferentes de los autores de la publicación;
  - no trata los datos como un ***registro de primera clase de investigación*** .
- El ***data paper*** como una solución.

CONCLUSIONES Y TRABAJOS FUTUROS

- La evaluación de métodos de extracción, data-cleaning y publicación en RDF del RDI DSpace y del SNDB accesibles a través de SPARQL endpoint públicos.
- Un prototipo de infraestructura con un conjunto de herramientas bien definidas, que resulta adecuada para los desarrollos futuros. Ej. Desarrollar ontologías propias del dominio que resulten necesarias para el RDI.
- Definición de una consulta integrada SPARQL que involucra dos o más conjuntos de datos RDF y que resuelve la vinculación de publicaciones científicas con datos primarios citados en ella.

CONICET

CENPAT

UNIVERSIDAD NACIONAL DE LA PATAGONIA SAN JUAN BOSCO

Referencias

- Altman, M., & King, G. (2008). A Proposed Standard for the Scholarly Citation of Quantitative Data. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1081955
- Arano, S., Martínez, G., Losada, M., Villegas, M., Casaldàliga, A., & Bel, N. (2011). La comunidad «Recursos y datos primarios» de la Universitat Pompeu Fabra: los repositorios institucionales como infraestructuras científicas: estudio de caso. Revista Española de Documentación Científica , 34 (3), 385–407.
- Ball, A., & Duke, M. (2011). How to cite datasets and link to publications. Digital Curation Centre .
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). THE SEMANTIC WEB. Scientific American , 284 (5), 34–43.

Referencias (cont.)

- Buneman, P., Davidson, S., & Frew, J. (2016). Why data citation is a computational problem. Communications of the ACM , 59 (9), 50–57.
- Davidson, S. B., Buneman, P., Deutch, D., Milo, T., & Silvello, G. (2017). Data Citation: a Computational Challenge. Proceedings of the ... ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems , 2017 , 1–4.
- Green, T. (2009). We need publishing standards for datasets and data tables. Learned Publishing: Journal of the Association of Learned and Professional Society Publishers , 22 (4), 325–327.
- Heath, T., & Bizer, C. (2011). Linked Data: Evolving the Web Into a Global Data Space . Morgan & Claypool Publishers.

Referencias (cont.)

- Hyland, B., Atemezing, G., & Villazón-Terrazas, B. (2014). Best practices for publishing linked data. W3C Working Group Note .
- Janowicz, K., Hitzler, P., Adams, B., Kolas, D., Vardeman, I. I., & Others. (2014). Five stars of linked data vocabulary use. Semantic Web , 5 (3), 173–176.
- Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: moving towards formal data publication. International Journal of Digital Curation , 6 (2), 4–37.
- Remsen, K. D. M. R. T. (2011). D, Braak. Darwin Core Archive How-To Guide .
- Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., … Desmet, P. (2014). The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. PloS One , 9 (8), e102623.

Referencias (cont.)

- Smith, M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., … Walker, J. H. (2003). DSpace: An Open Source Dynamic Digital Repository. DSpace: An Open Source Dynamic Digital Repository. Smith, MacKenzie; D-Lib Magazine , 9 (1). https://doi.org/ 10.1045/january2003-smith
- Starr, J., & Gastl, A. (2011). isCitedBy: A Metadata Scheme for DataCite. Retrieved from https://escholarship.org/uc/item/6r03h784
- Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., …
- Vieglais, D. (2012). Darwin Core: an evolving community-developed biodiversity data standard. PloS One , 7 (1), e29715.
- Zárate, M., Braun, G. A., & Fillottrani, P. R. (2017). Adding Biodiversity Datasets from Argentinian Patagonia to the Web of Data. In International Semantic Web Conference (Posters, Demos & Industry Tracks). Retrieved from http://ceur-ws.org/Vol-1933/paper-6.pdf