Datos Enlazados y Arrays

José R.R. Viqueira

Centro Singular de Investigación en Tecnoloxías da Información (CITIUS)
Rúa de Jenaro de la Fuente Domínguez,
15782 - Santiago de Compostela.
Despacho: 209
Telf: 881816463
Mail: jrr.viqueira@usc.es
Skype: jrviqueira
URL: http://citius.usc.es/equipo/persoal-adscrito/jrr.viqueira

Curso 2021/2022

Guion

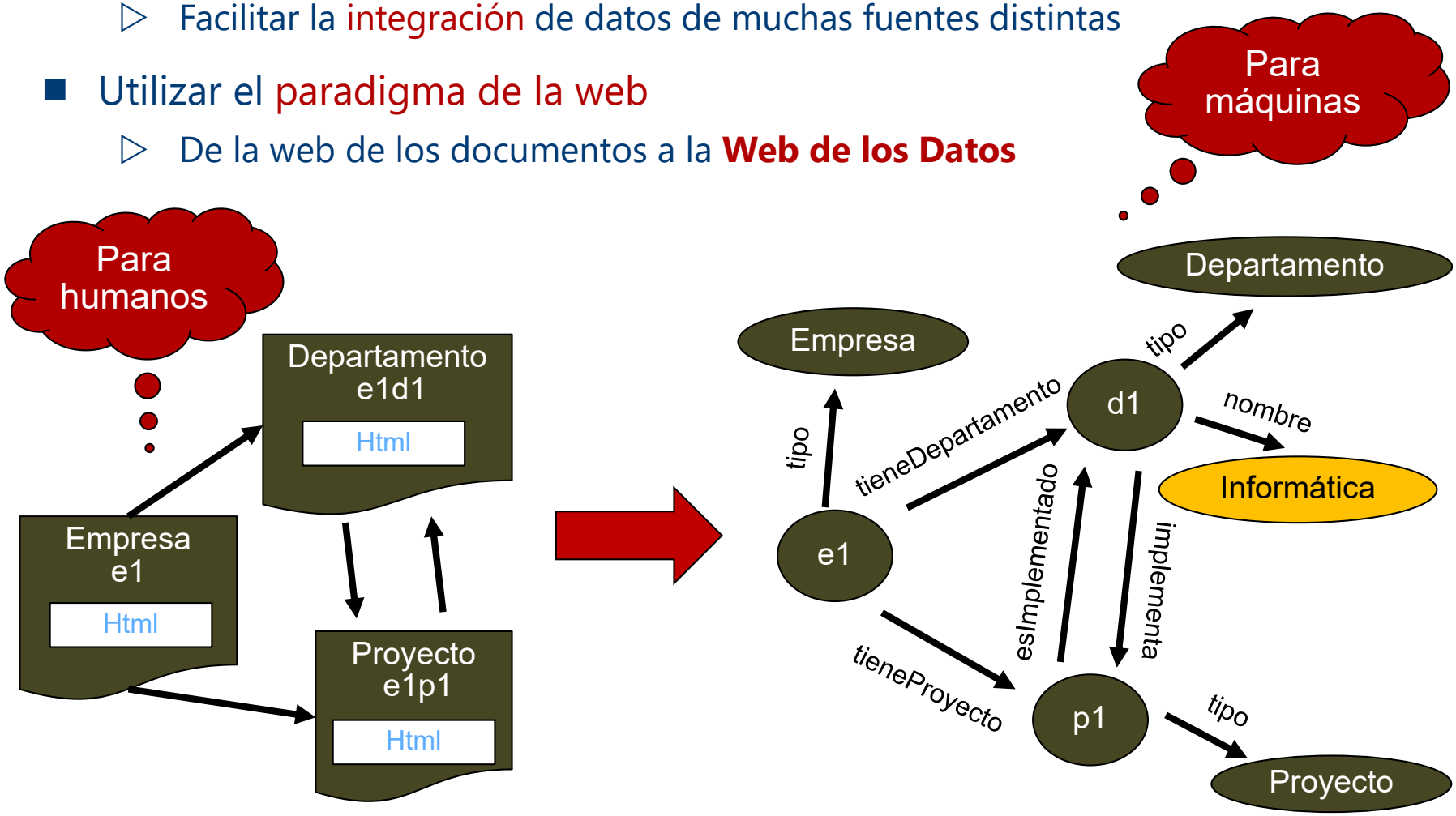
- **Datos Enlazados**
  - Introducción
  - Principios
  - RDF y SPARQL
  - Conclusiones
- **Almacenamiento y gestión de arrays**
  - Archivos
  - Bases de datos objeto-relacionales
  - Bases de datos de arrays

Datos Enlazados: Introducción

- Diluvio de datos (**Data Deluge**): Multitud de conjuntos de datos generándose cada día.
- Islas de datos
  - Servicios web independientes
  - Formatos XML, JSON, CSV, etc.
  - Problemas
    - Uso de identificadores locales en cada fuente de datos
      - Código: dinf, Nombre: 'Departamento de informática'
    - No existen relaciones entre datos de distintas fuentes (Relaciones externas)
  - Los datos de cada conjunto de datos están aislados de los demás y no se pueden
    - Descubrir unos navegando desde los otros (como en la web)
    - Combinar de forma fácil para utilizar de forma conjunta
      - Problemas de semántica en la integración

Datos Enlazados: Introducción

- Necesidades
  - Tecnologías de acceso que faciliten la reutilización de los datos
  - Facilitar el descubrimiento de los datos
  - Facilitar la integración de datos de muchas fuentes distintas
- Utilizar el paradigma de la web
  - De la web de los documentos a la **Web de los Datos**



Datos Enlazados: Principios

Propuestos por Tim Berners Lee

1. **Uso de Universal Resource Identifier (URI) como nombre de cosas**
   - ▷ Forma simple para crear identificadores globales
     - _ http://dbpedia.org/resource/Santiago_de_Compostela
2. **Uso de URIs HTTP para proporcionar acceso a la información sobre las cosas**
   - ▷ Habilita la posibilidad de implementar herramientas: Navegadores, Buscadores.
3. **Utilización de estándares (RDF, SPARQL) para representar y consultar información sobre las cosas**
4. **Incluir enlaces desde una cosa a las otras, para poder descubrir nuevas cosas a partir de la información ya accedida.**
   - ▷ Enlaces externos son fundamentales para la Web de los Datos
   - ▷ Tipos de enlaces
     - _ Relaciones entre cosas
     - _ Enlaces de identidad (owl:sameAs)
       - ▪ Apuntan a representaciones distintas de la misma cosa en fuentes distintas
     - _ Enlaces de vocabulario (rdf:type)
       - ▪ Definiciones de términos del vocabulario utilizado.
       - ▪ Datos autodescriptivos. Integración de datos a través de vocabularios.

Datos Enlazados: RDF y SPARQL

Resource Description Framework (RDF)

- https://www.w3.org/TR/rdf-concepts/
- Modelo de datos estandarizado para la Web de los Datos
  - Representación de la información como un grafo dirigido con arcos etiquetados
  - Uso como modelo común para la integración de fuentes heterogéneas
  - Recursos descritos con conjuntos de tripletes (sujeto, predicado, objeto)
    - Sofia tiene edad 32
  - Sujeto: URI que identifica al recurso que se está describiendo
  - Predicado: Tipo de relación que existe entre el sujeto y el objeto. Se identifica también con una URI. Viene de un vocabulario
  - Objeto
    - Literal: Texto con información del lenguaje. Valor con tipo de datos.
    - URI: Para proporcionar enlaces entre recursos.
- Serializaciones de los datos entendibles por humanos y procesables por máquinas.
  - RDF/XML
  - Turtle
  - JSON-LD

Datos Enlazados: RDF y SPARQL

- **SPARQL Protocol and RDF Query Language (SPARQL)**
  - Lenguaje estandarizado para consultar fuentes RDF

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select STR(?nprov) as ?provincia STR(?nmun) as ?municipio
where {
 <http://dbpedia.org/resource/Galicia_(Spain)> dbo:subdivision ?prov.
 ?prov rdfs:label ?nprov.
 ?prov dbo:type dbr:Provinces_of_Spain.
 ?mun dbo:subdivision ?prov.
 ?mun rdfs:label ?nmun.
 FILTER (LANG(?nmun) = "es" and LANG(?nprov)="es")
}
order by ?nprov, ?nmun
```

Datos Enlazados: Conclusiones

- **Ventajas** de usar Datos Enlazados para publicar en lugar de habilitar la descarga en formato convencionales (CSV, JSON, XML, etc.)
  - Facilita procesos de descubrimiento y consulta sobre fuentes de datos distintas.
  - Más en concreto:
    - Proporciona un modelo de datos unificado: RDF
    - Proporciona una forma de acceso a datos estandarizada basada en HTTP.
      - Evita interfaces propietarios.
    - Habilita el descubrimiento de datos basado en enlaces, en una Web de los Datos de ámbito global.
      - Uso de URIs como identificadores globales.
      - Uso de URIs para enlazar descripciones de recursos en fuentes distintas
    - Los datos pueden autodescribirse utilizando vocabularios compartidos entre fuentes distintas.

Datos Enlazados: Conclusiones

- Esquema de desarrollo progresivo de la Web de los Datos propuesto por Tim Berners Lee

★ Datos publicados en la web con una licencia Open Data

★★ Además, uso de formatos de datos estructurados apropiados para su procesamiento por máquinas (ejemplo: MS Excel)

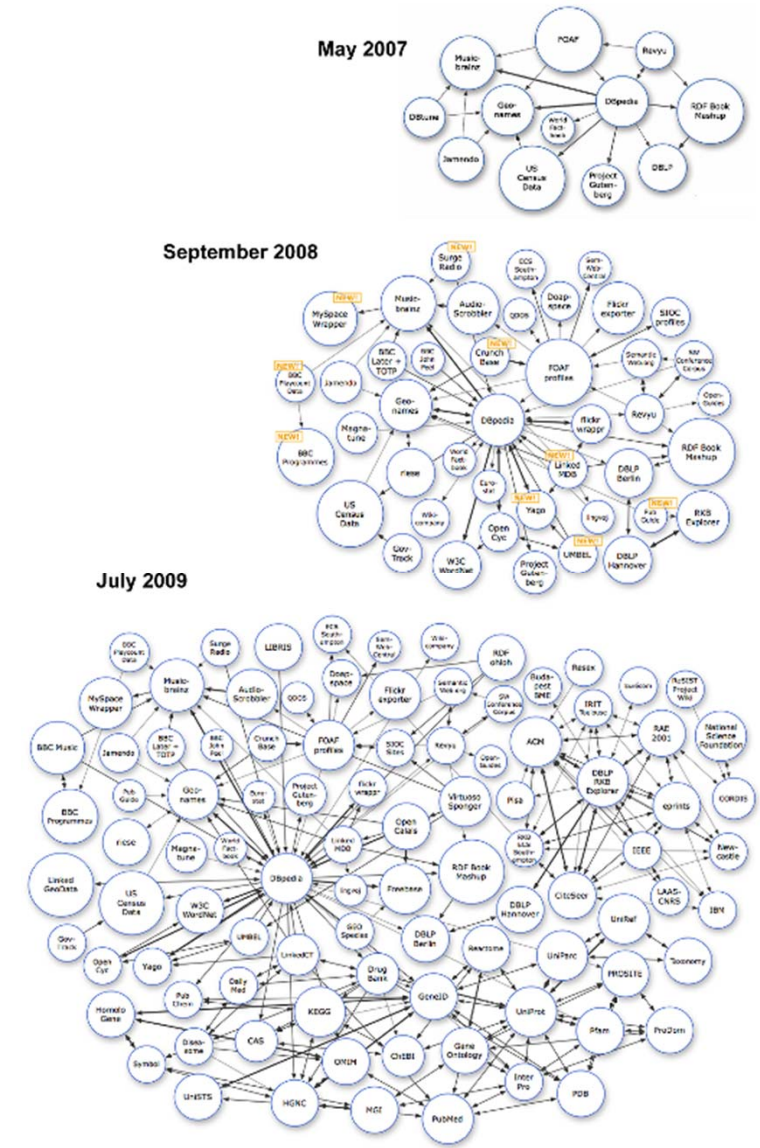
★★★ Además, uso de formatos no propietarios para los datos (ejemplo: CSV, XML, JSON, etc.)

★★★★ Además, uso de estándares del WWW Consortium (W3C) del ámbito de la web semántica representar y acceder a los datos (RDF, SPARQL).

★★★★★ Además, proporcionar enlaces a otras fuentes de datos y vocabularios cuando sea necesario identificar el contexto de los datos publicados.

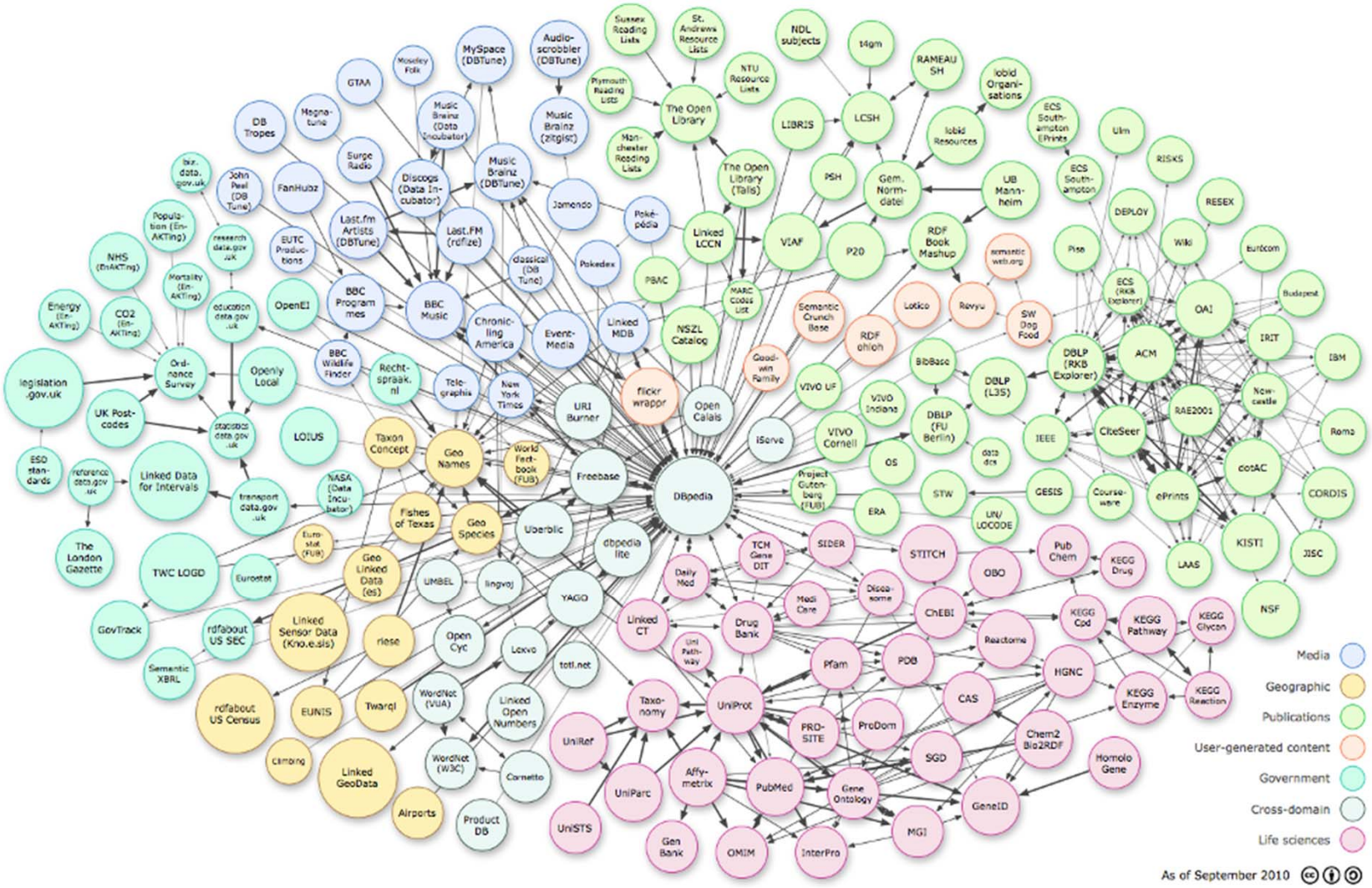
Datos Enlazados: Conclusiones

- Evolución



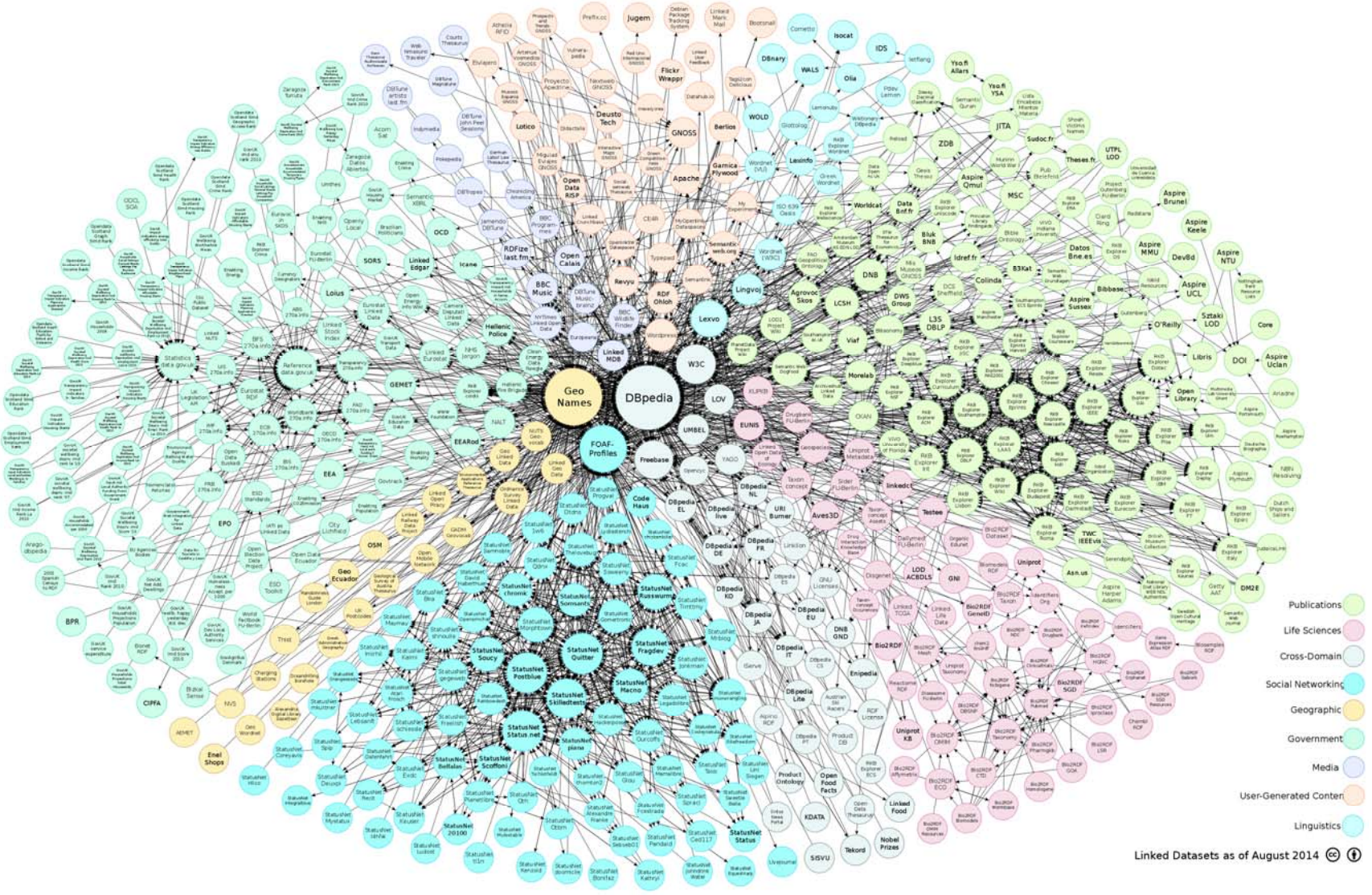
Datos Enlazados: Conclusiones

- Evolución



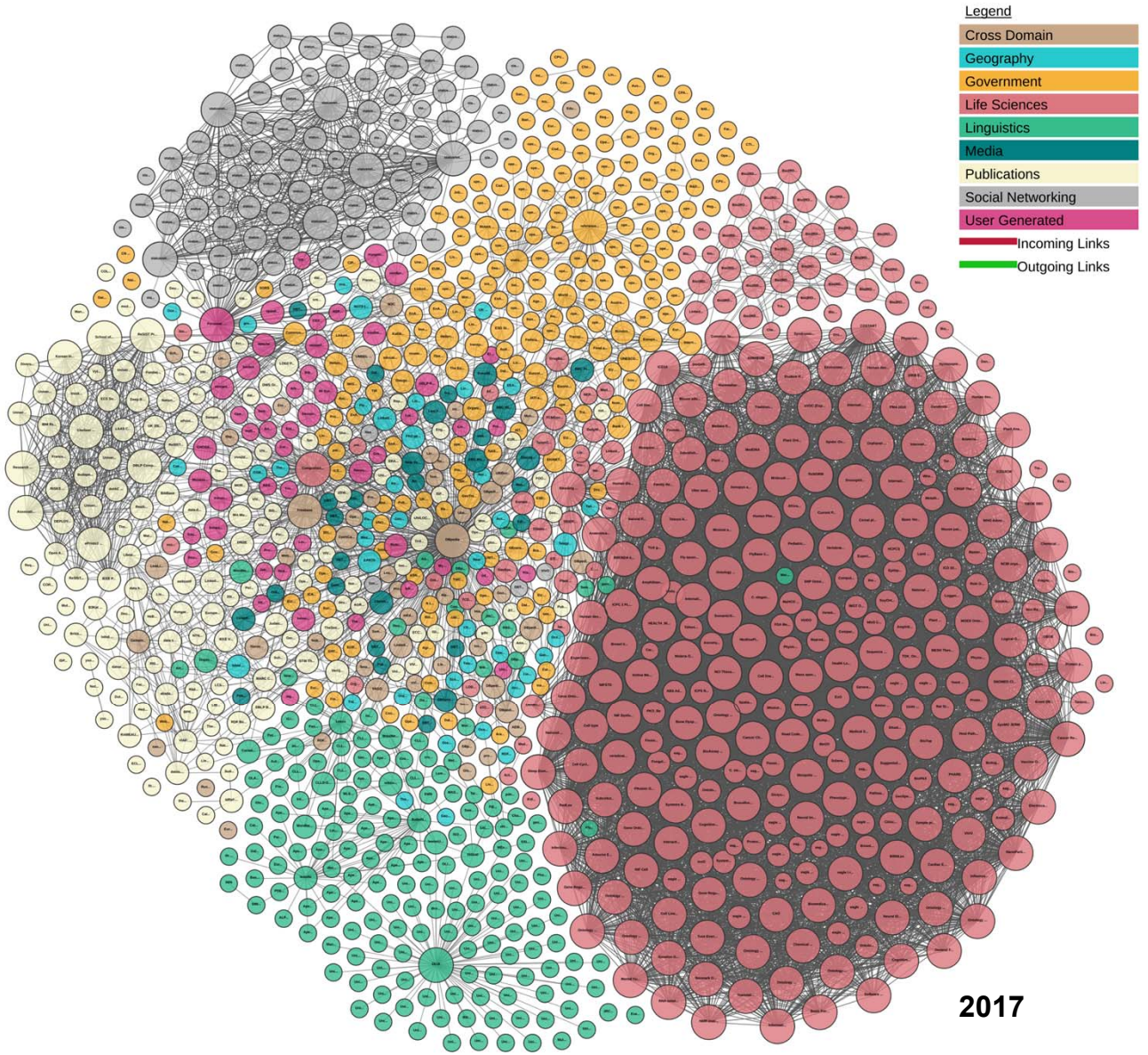
Datos Enlazados: Conclusiones

- Evolución



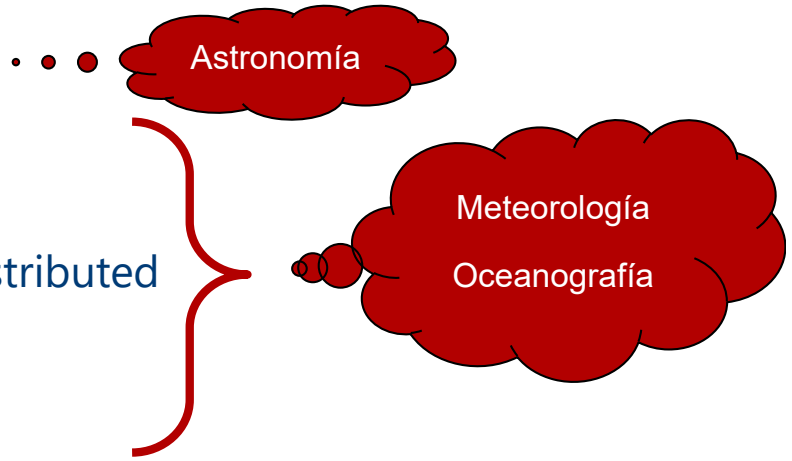
Datos Enlazados: Conclusiones

- Evolución



Almacenamiento y gestión de Arrays: Archivos

- ¿Por qué no se gestionan de forma eficiente un modelo de datos convencional?
  - No es necesario almacenar siempre los valores de las dimensiones
  - El orden de los datos es muy importante
- Formatos para imagen geográfica
  - GeoTIFF (Formato TIFF con cabecera geográfica)
  - ECW (Enhanced Compressed Wavelet)
- Arrays e imágenes en entornos geocientíficos (Ejemplos)
  - Flexible Image Transport System (FITS)
  - Hierarchical Data Format (HDF4, HDF5)
  - GRIdded Binary or General Regularly-distributed Information in Binary form (GRIB)
  - Network Common Data Form (NetCDF)



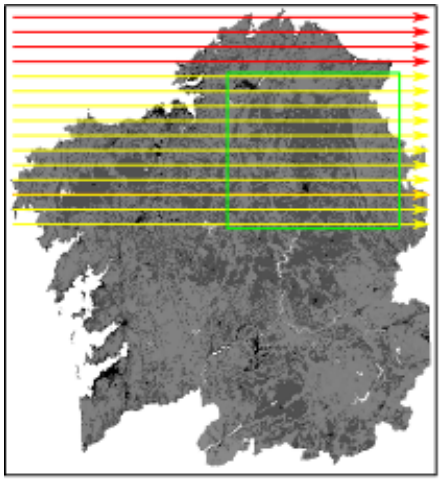
Almacenamiento y gestión de Arrays: Bases de datos Objeto-Relacionales

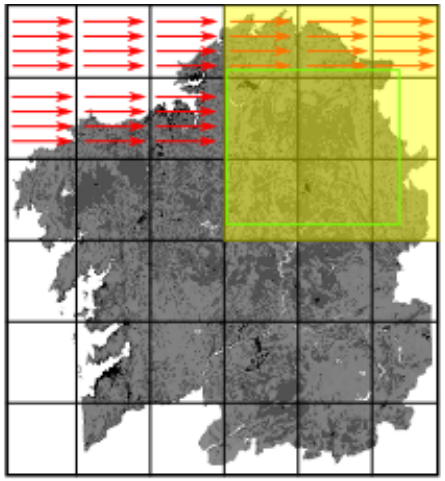
- Constructor de tipos ARRAY aparece en el estándar SQL:2003
- Sintaxis y funciones para
  - Acceder a elementos por índice
  - Desanidar arrays, materializando si es necesario el orden de los elementos
  - Anidar elementos en arrays en un determinado orden
- Limitaciones
  - El estándar se limita a arrays de una sola dimensión
  - No está diseñado para gestionar arrays de grandes dimensiones (orden de los Terabytes o Petabytes) como los que aparecen en aplicaciones del ámbito científico.

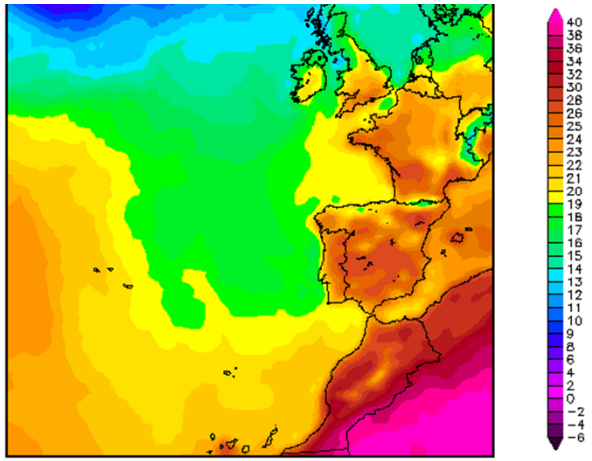
```
CREATE TABLE Empleado (
  id_emp INTEGER PRIMARY KEY,
  nombre VARCHAR(50),
  salarios DECIMAL(8,2) ARRAY[12]
)
```

Almacenamiento y gestión de Arrays: Bases de datos Objeto-Relacionales

- Extensión para ARRAYS geoespaciales (RASTER)
  - Disponible en extensiones espaciales de BD OR
    - Ejemplos: Oracle Spatial, PostGIS para PostgreSQL
  - Tipo de datos RASTER
    - Combina un array bidimensional con metadatos para georreferencia en la superficie terrestre
  - Funciones y operadores para
    - Elementos de tipo raster
    - Elementos de tipo geométrico (puntos, líneas, polígonos, ...)
  - Combinación con tipos temporales
  - Problemas Principales
    - Eficiencia en grandes arrays
    - Usuario debe gestionar las teselas





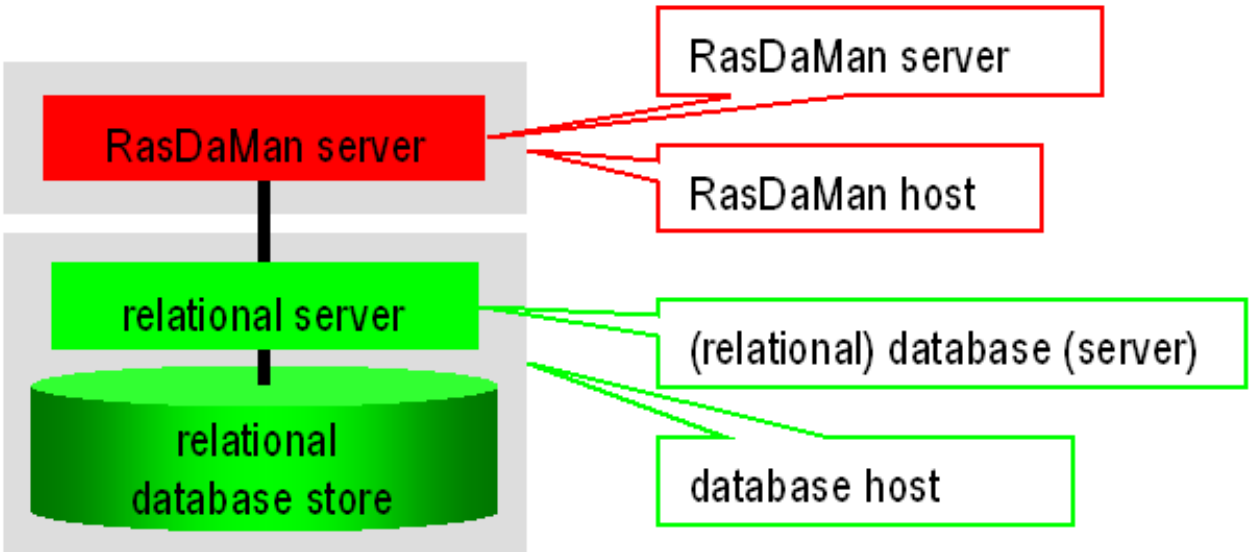



```
CREATE TABLE MDT (
  id_tesela INTEGER PRIMARY KEY,
  tesela RASTER)
```

```
CREATE TABLE meteo (
  id_tesela INTEGER PRIMARY KEY,
  tiempo TIMESTAMP,
  tesela RASTER)
```

Almacenamiento y gestión de Arrays: Bases de datos de Arrays

- Rasdaman (http://www.rasdaman.org/)
  - Gestor de arrays multidimensionales
  - Implementado sobre un SGBDs relacional
    - Típicamente PostgreSQL
    - Arrays almacenados en tipos BLOB



Almacenamiento y gestión de Arrays: Bases de datos de Arrays

- Rasdaman (http://www.rasdaman.org/)
  - Lenguaje de consulta declarativo sobre arrays (rasql)
  - Lenguaje de definición de arrays (rasdl)
    - Definición del tipo de las celdas del array
      - Tipos simples y tipos complejos (struct)
    - Definición de arrays sobre un tipo de celda
    - Definición de colecciones de arrays del mismo tipo
  - Lenguaje de manipulación de arrays (rasml)

```
select RGBSet
from RGBSet
where
 all_cells(RGBSet.green>20)
```

all_cells es una función de agregado que devuelve true si todas las celdas devuelven true

Arrays en los que todos elementos del **green** son mayores que 20

```
struct RGBPixel{char red, green, blue;};

typedef marray <RGBPixel,[0:799, 0:599]> RGBImage;

typedef set <RGBImage> RGBSet;
```

```
select RGBSet[120:160, 55:75]
from RGBSet
```

```
select png(RGBSet)
from RGBSet
```

```
select
    RGBSet.red * 2
from RGBSet
```

Crea histogramas para el atributo green

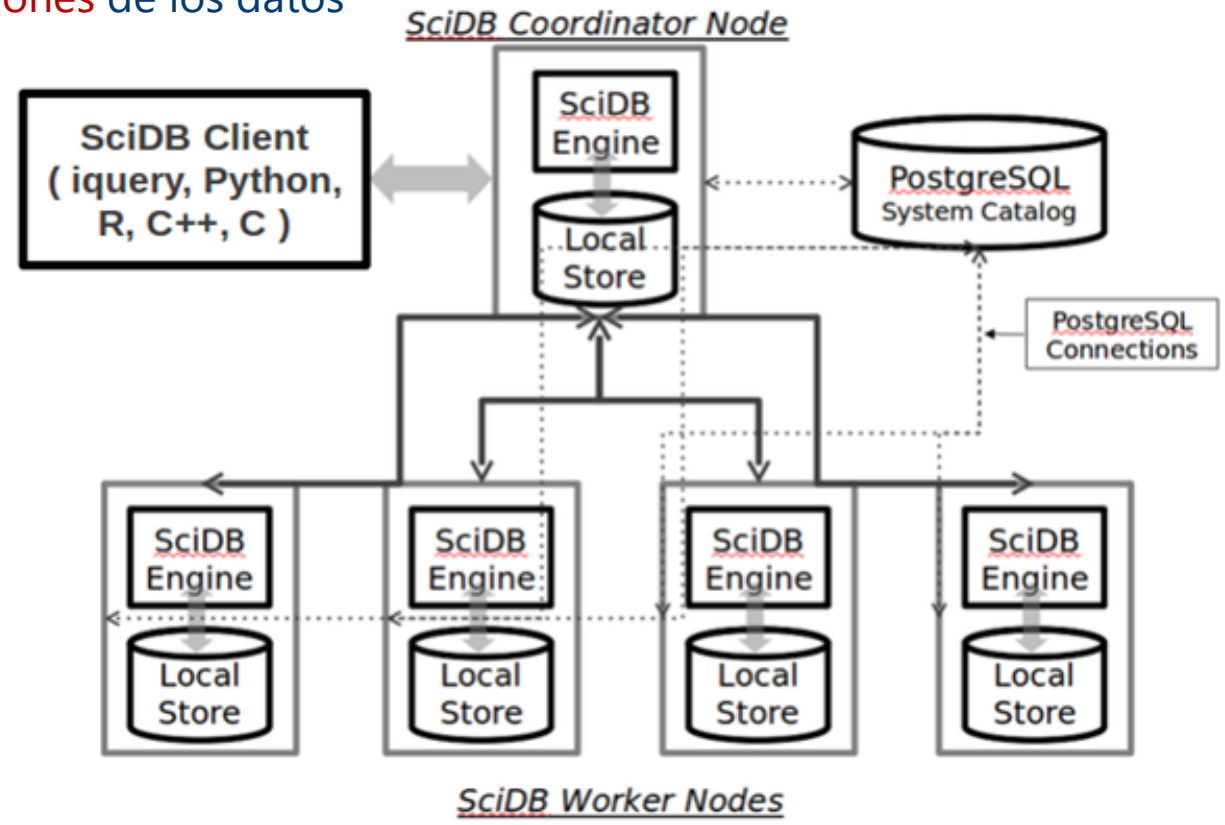
Crea un nuevo array **v** de 256 elementos por cada array **r** en RGBSet

Cada elemento de **v** tiene el número de celdas con ese valor en el atributo green en **r**.

```
select marray v in [0:255]
       values condense +
              over x in sdom(RGBSet)
              where RGBSet[x].green = v[0]
              using 1
from RGBSet
```

Almacenamiento y gestión de Arrays: Bases de datos de Arrays

- SciDB (https://www.paradigm4.com/)
  - Operadores para Arrays y Vectores
  - Arquitectura paralela de tipo Share-Nothing
  - Extensible: Tipos de dato y funciones definidos por el usuario
  - Integración con R
  - Gestión de versiones de los datos

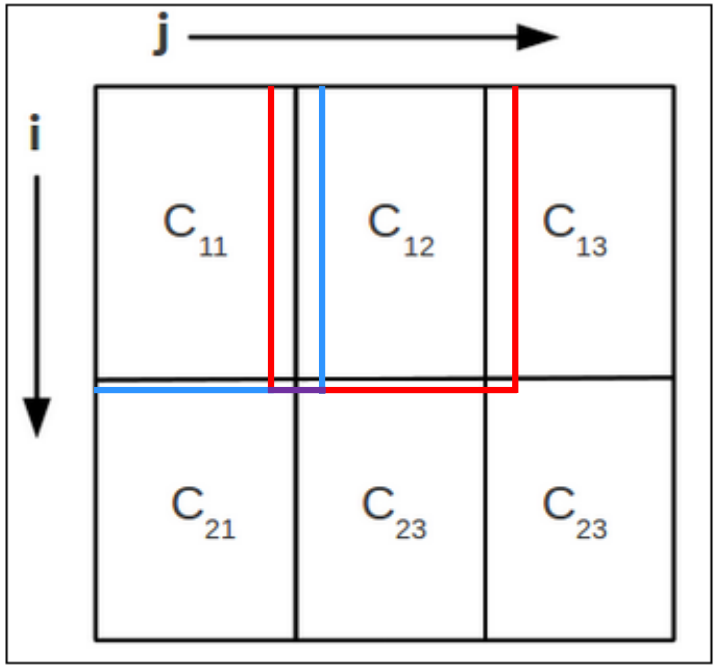


Almacenamiento y gestión de Arrays: Bases de datos de Arrays

- SciDB (https://www.paradigm4.com/)
  - Modelo de datos de Arrays (nombre, dimensiones, atributos)
  - Chunking
    - Cada dimensión dividida en chunks (X chunks de tamaño Y)
    - Distribución de los chunks en nodos usando hashing
    - Solapamiento de los chuncks
      - A <a: int32> [i=1:10,5,1, j=1:30,10,5]
      - Mejora rendimiento en operaciones de vecindario

2 chunks de tamaño 5 con solapamiento de 1 en la dimensión i

3 chunks de tamaño 10 y solapamiento 5 en la dimensión j

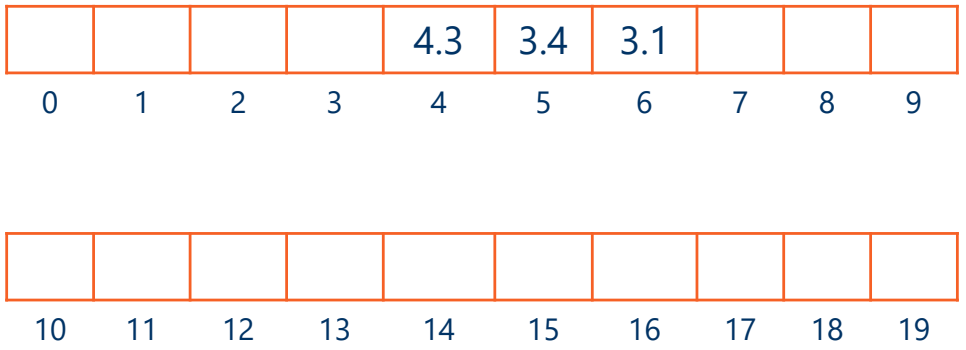


Almacenamiento y gestión de Arrays: Bases de datos de Arrays

- SciDB (https://www.paradigm4.com/)
  - Lenguajes de consulta y análisis
    - Array Query Language (AQL)
    - Array Functional Language (AFL)

```
CREATE ARRAY A
   <val_a:double>[i=0:19,10,0];

SELECT sqrt(val_a)
FROM A
WHERE i>3 AND i<7
```



Datos Enlazados y Arrays

José R.R. Viqueira

Centro Singular de Investigación en Tecnoloxías da Información (CITIUS)
Rúa de Jenaro de la Fuente Domínguez,
15782 - Santiago de Compostela.
Despacho: 209
Telf: 881816463
Mail: jrr.viqueira@usc.es
Skype: jrviqueira
URL: http://citius.usc.es/equipo/persoal-adscrito/jrr.viqueira

Curso 2021/2022