Módulo IV
Acceso y recuperación de datos en la Web

OpenCourseWare

Recuperación y Acceso a la Información

Contenidos

- Ciclo de Vida de la Información
- Fuentes de la Información
- Estructuración y Saneamiento de los datos: la coherencia
- Integración de fuentes

De los datos a la Información

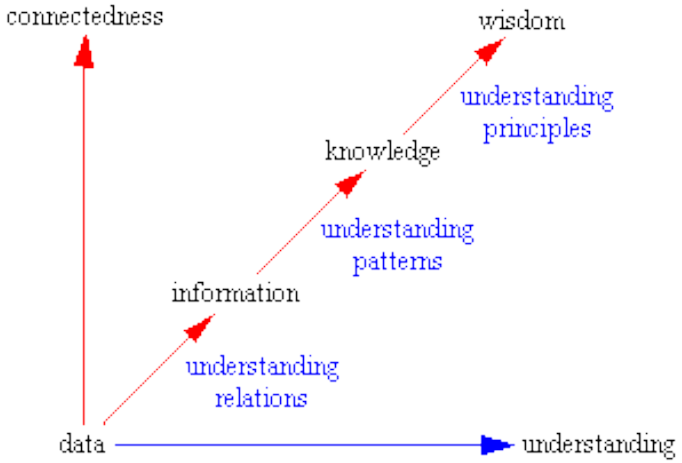
La gestión de información comporta aumentar relaciones y clasificaciones para mejorar su comprensión y los procesos asociados.

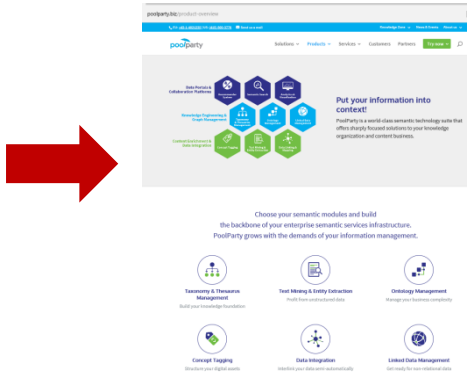
Datos. Símbolos o signos en relación a una diferencia observada

Información. Relaciones entre datos en un dominio, para resumir , clasificar, etc

Conocimiento. Información almacenada con un objetivo para inferir nueva información basada en un patrón, know-how

Sabiduría: teorías, principios… es decir patrones de conocimiento





Ejemplo Pool-party

La gestión de información se puede sistematizar en Ciclos de Vida

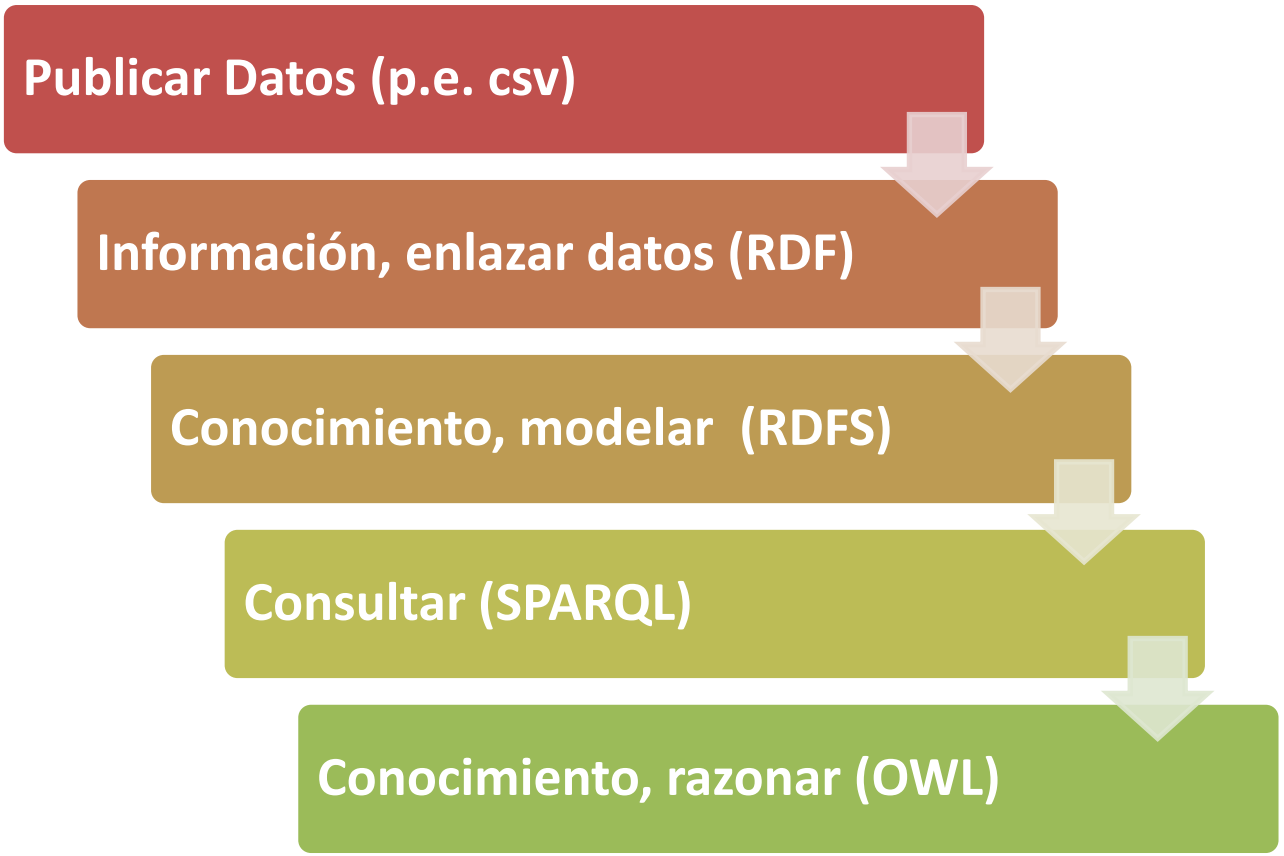
Ciclos de Vida

Minería de datos

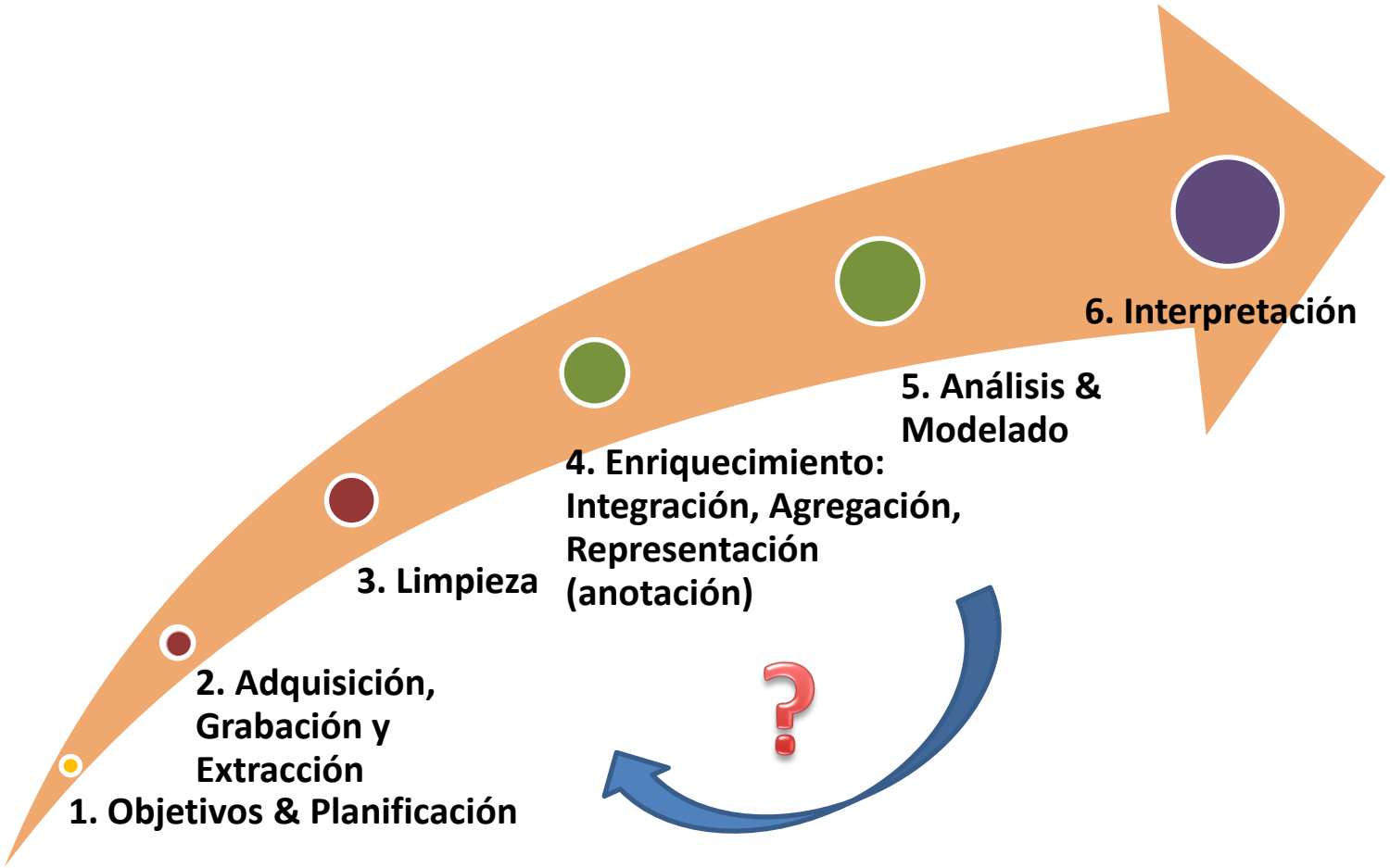
Desarrollo de Software

Ontologías

Proceso en la Web Semántica



Ciclo de vida de los datos web



Por ejemplo, Big Data

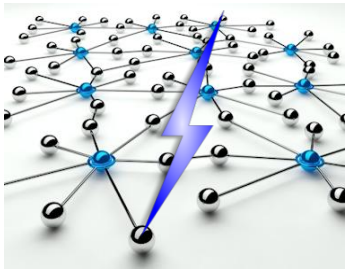
Qué es Big Data

Tecnología y capacidad del manejo de datos, debido a problemas con:

Capacidad Tecnológica (3V)

Volumen

+



Velocidad

+

Variabilidad

Conclusiones válidas (4V)

+

Veracidad

Fuente: IBM, Forbes, Experian QAS, Gardner

Etapas del ciclo de vida en Big Data

55% proyectos fallidos por mala planificación

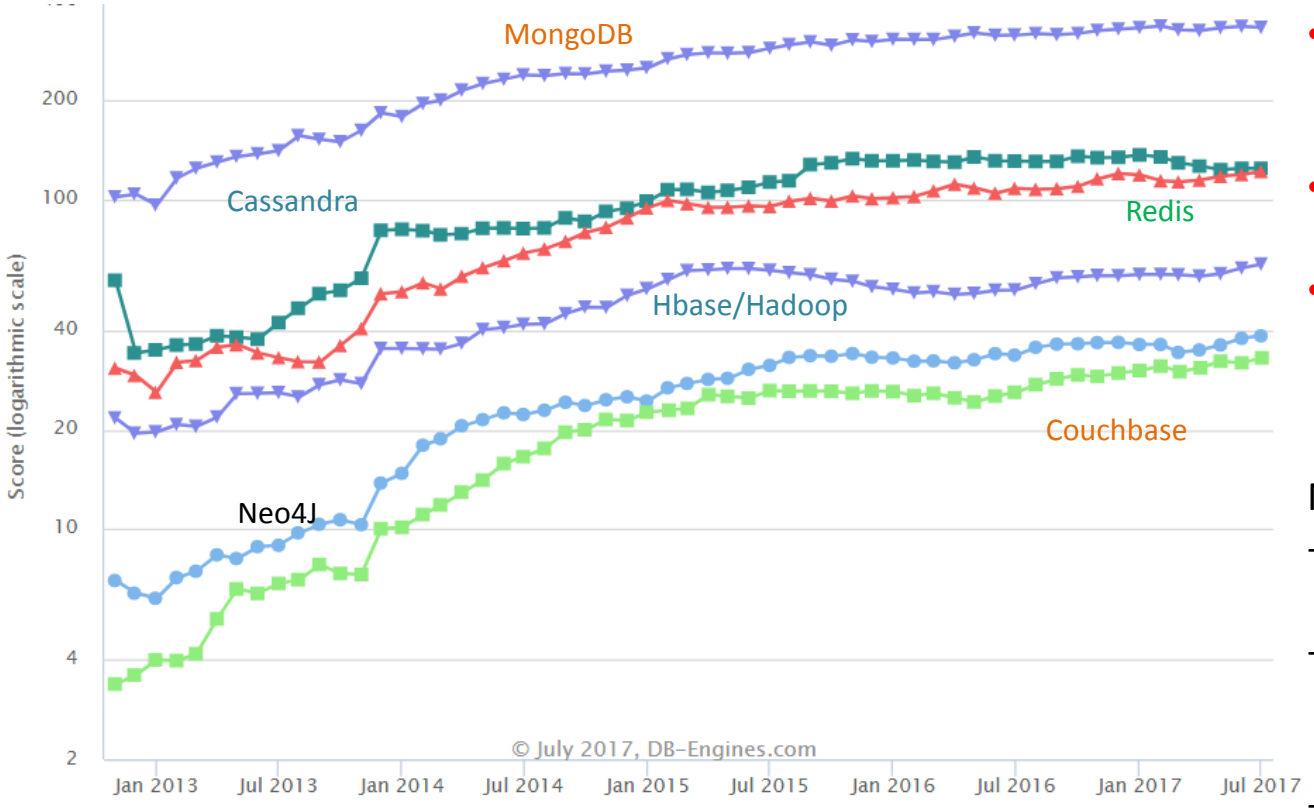
80% Coste del proyecto

- **Objetivos y planificación**: que datos tenemos y datos esperamos obtener.
- **Adquisición**: adquirir información de fuentes internas o **externas**.
  Utilizar fuentes reputadas y registrarlo (data lineage)
- **Limpieza**: eliminar **inconsistencias y datos corruptos**.
- **Integración**: enriquecimiento de datos mediante su **integración coherente**, para dar mayor **contexto** y significado.
- **Análisis**: Modelado de datos y análisis para **responder preguntas** sobre una organización.
- **Interpretación**: propuesta de unas **pautas para facilitar la gestión** de los datos

Fuente: http://db-engines.com/en/ranking

Documental
Grafo
Atributo-valor
Columnar

NoSQL BD



NoSQL:

- *alta disponibilidad,*
- *consistencia eventual, sin licencias,*
- *web (operaciones simples y tiempo real)*
- *Join inexistententes o limitados*

No garantizan **ACID**

- *Atomicidad* transacción completa. Todo o nada
- *Consistencia:* tras la transacción las restricciones de la BD se cumplen
- *Isolation* las transacciones son secuenciales
- *Durabilidad* las transacciones no se pierden aunque caiga el sistema

No garantizan **CAP**

- **Consistencia**: cualquier nodo mismo resultado
- **Disponibilidad**: si un nodo recibe una petición siempre hay respuesta
- **Tolerancia a Particiones**: el sistema responde aunque falle algún nodo

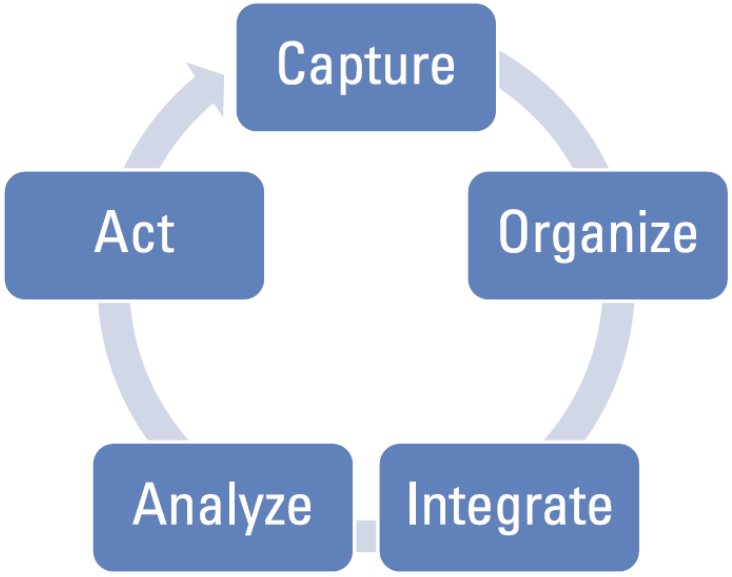
AP (Cassandra),
CP (Mongo),
CA (BDR)

Contenido

- Ciclo de Vida de la Información
- Fuentes de la Información
- Estructuración y Saneamiento de los datos: la coherencia
- Integración de fuentes

Planificación

- Plantear bien la cuestión que se quiere responder: datos de partida y datos que pretenden conseguir
- Identificar el data set idóneo
- Determinar las fuentes de información apropiadas

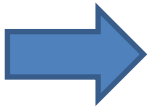


Planificación: Calidad

Objetivo: identificar la fuente idónea teniendo en cuenta su utilidad para nuestro problema, veracidad, calidad y formato.

Planificación:

1. Identifica el problema
2. Comienza con una muestra
3. Realizar un caso de estudio
4. Identificar necesidades de recolección, limpieza y agregación



Datos

- Completos
- Tipología válida
- Veraces
- Precisos
- Sin duplicados
- Procedencia conocida
- Actualizados
- Disponibles
- Relevantes
- Fiables
- Consistentes

Planificación: Calidad

- **Calidad**: consistente (coherente y uniforme), completa y limpia
- **Integridad de datos** (sin perdida de datos, incorrectos, contaminados, inconsistentes o duplicados)
- **Novedad**: la afirmación es nueva para mí
- **Validez**: fuente fiable, con autoridad y con información sobre su procedencia y como se ha transformado (data lineage).
- **Veracidad** tiene tres dimensiones: objetividad, veracidad, credibilidad
- **Utilidad**: si tiene un impacto sobre el objetivo de mi proyecto

Si se siguen normas de calidad como normalizar terminología, asegurar interoperabilidad y asegurar integridad de datos y data lineage, el coste de limpieza disminuirá

Las unidades de medida y la descripción suelen venir en el documento ***Codebook***
El cómo se adquiere y se transforma se registra en el ***Diseño de Estudio***

Adquisición y limpieza

- Raw data: datos en bruto, tal cual los encontramos
  - Mediante procesamiento se transforman en datos limpios
  - Procesamiento: descripción, segmentación, normalización, corrección, fusión y transformación.
- Adquisición: identificar e importar los datos
- Limpieza (data cleaning, cleansing o scrubbing): identificar registros erróneos, no fiables o inconsistentes para mejorar su calidad
- La calidad suele estar relacionada con la procedencia, obtención y modificación (***data lineage***)

Adquisición: Fuentes de Información

Existen múltiples fuentes que suelen integrarse:

Procedencia

- Datos públicos: muchos con estructuración pobre
- Datos internos de la organización

Creación

- Manual
- Automática: p.e. con redes de sensores

Formalización

- No estructurados (80% de los datos de las empresas), p.e. Lenguaje Natural
- Semiestructurados: XML
- Estructurados: tipo dato conocido, con esquema y restricciones de datos

Adquisición: Acceder a los datos

Existe un gran número de datasets públicos multitemáticos:

US Gob. DataSources	http://usgovxml.com/
Amazon (necesidad de validarse y ejecutar una instancia EC2, EBS)	http://aws.amazon.com/es/public-data-sets/
Datalib	http://databib.org
Datacite	http://datacite.org
Figshare	http://figshare.com
Linked data	http://linkeddata.org
DataHub	http://thedatahub.org
Enigma (buscador)	https://app.enigma.io/
Quandl (buscador)	https://www.quandl.com/
Google Public Explorer	https://www.google.com/publicdata/directory
Public Data commons (hasta 1PB)	https://www.opensciencedatacloud.org/publicdata/

Formatos de serialización

Traducción de estructuras de datos en formato de texto o binario para su tranferencia o almacenamiento

	Pros	Cons
XML	texto plano, legible	Tamaño del fichero
Binary XML	Ligero en cuanto tamaño	No texto plano
JSON	Ligero, texto plano, legible	Pocos tipos de datos
YAML	Compacto y legible	Pocos tipos de datos
RDF	texto plano, estructurado, legible	Pocos tipos de datos
BSON	Eficiente y con tipos de datos	No legible
CSV	Ligero y legible	Pobre en tipos de datos y estructura

Formatos de serialización

Formatos	Ejemplo
TSV	NAME NACIONALITY WEIGHT Alan Spanish 55 John French 129
CSV	NAME, NACIONALITY,WEIGHT Alan, Spanish, 55 John, French, 129
XML	<example> <person ID="1"> <name> Alan </name> <nationality> Spanish</nationality><weight>55</weight></person> <person ID="2"><name>John<….></person></example>

Formatos de serialización

Formatos	Ejemplo
JSON	{“example":[{“name":“Alan”, “nacionality”:“Spanish“, “phone“:[“work_ph”:”25255”,”cell_ph”:”45433”] , “weight”:51}, {other_record}]}
JSON-based	BSON (Binary JSON) basado en JSON, con tipos de datos (string, Integer, double, fecha, array o booleano), tamaño del documento y longitud del campo. Otras basadas en JSON son: HOCON, Candle , Smile or Yaml
YAML	Data: given: Alan nacionality: Spanish weight: 51.5 age: 26 Phone: - Work: 25255 Address: 8 St.Paul Av. Quebec - cellular: 45433

Formatos de serialización: LoD

Formatos	Ejemplo
RDF	<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns: foaf: “http://xmlns.com/foaf/” xmlns:ex=“http://example.edu/ ”> <rdf:Description rdf:about="http://example.edu/alan"> <rdf:type rdf:resource:”http://xmlns.com/foaf/person” /> <ex:name >Alan</ex:name> <ex:weight>57</ex:weight> </rdf:Description> </rdf:RDF>
TURTLE	PREFIX ex: <http://example.edu/> PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns> PREFIX foaf: <http://xmlns.com/foaf/> <ex:alan> a foaf:Person ; <ex:name>Alan <ex:weight>57

Contenido

- Ciclo de Vida de la Información
- Fuentes de la Información
- Estructuración y Saneamiento de los datos: la coherencia
- Integración de fuentes

Limpieza

Tareas encaminadas a la eliminación de errores ortográficos y tipográficos, obtener datos normalizados y validar su corrección

Algunas estrategias :

- Distancias de edición (p.e. distancia Levehenstein)
- Comparadores fonéticos (p.e. Soundex y Metaphone)
- Frecuencia de subcadenas (p.e. ngram y fingerprint)
- Crowdsourcing
- Comparación con diccionarios y listados
- Visualización y distribuciones estadísticas anómalas

Beneficios:

✓Posibilitar un análisis estadístico fiable

✓ Capacidad de integración con otros datasets

✓Calidad de datos: validar los datos localmente y compararlos con fuentes externas

Tareas para realizar la limpieza de datos

- Parsing / Extraer las cadenas alfanuméricas
- Transformación de:
  - Múltiples representaciones con la misma semántica
  - Datos no atómicos (separar los elementos: split)
  - Datos no estructurados
  - Identificar errores tipográficos y ortográficos
  - Datos inconsistentes (no coherentes y uniformes)
  - Datos incompletos o ambiguos
- Eliminación de duplicados

Planificación

Evitar reevaluar datos ya limpiados y sistematizar y documentar el flujo de trabajo (workflow) para optimizar la limpieza de nuevos datos.

Limpieza de datos. Ejemplo

"Juan Garcia, Av. Pez,8, Mejico DF, 40205 México"
"J. García, Avenida Pez,8, México DF, 50205 México"

Proceso	Ejemplo	Resultado ejemplo
Split	Juan Garcia, Av. Pez,8, Mejico DF, 40205 México	Juan; Garcia; Av. Pez 8; Mejico DF; 40205; México
Typos	Juan; Garcia; Av. Pez 8; Mejico DF; 40205; México	Juan; García; Av. Pez 8; Méjico DF; 40205; México
Normalizar	Juan; García; Av. Pez 8; Méjico DF; 40205; México	Juan; García; Av. Pez 8; México DF; 40205; México
Validar	Juan; García; Av. Pez 8; México DF; 40205; México	Juan; García; Av. Pez 8; México DF; 50205; México
Agrupar	Juan García, Av. Pez,8, México DF, 50205 México J. García, Avenida Pez,8, México DF, 50205 México	Juan García, Av. Pez,8, México DF, 50205 México Juan García, Av. Pez,8, México DF, 50205 México

Variantes en Google SE con Britney Spears

Distintas variaciones con más de 40 ocurrencias de las 593 registradas con frecuencia mayor de dos:

488941 britney spears
40134 brittany spears
36315 brittney spears
24342 britany spears
7331 britny spears
6633 briteny spears
2696 britteny spears
1807 briney spears
1635 brittny spears
1479 brintey spears
1479 britanny spears
1338 britiny spears
1211 britnet spears
1096 britiney spears
991 britaney spears
991 britnay spears
811 brithney spears
811 brtiney spears

664 birtney spears
664 brintney spears
664 briteney spears
601 bitney spears
601 brinty spears
544 brittaney spears
544 brittnay spears
364 britey spears
364 brittiny spears
329 brtney spears
269 bretney spears
269 britneys spears
244 britne spears
244 brytney spears
220 breatney spears
220 britiany spears
199 britnney spears
163 britnry spears

147 breatny spears
147 brittiney spears
147 britty spears
147 brotney spears
147 brutney spears
133 britteney spears
133 briyney spears
121 bittany spears
121 bridney spears
121 britainy spears
121 britmey spears
109 brietney spears
109 brithny spears
109 britni spears
109 brittant spears
98 bittney spears
98 brithey spears
98 brittiany spears

98 btitney spears
89 brietny spears
89 brinety spears
89 brintny spears
89 britnie spears
89 brittey spears
89 brittnet spears
89 brity spears
89 ritney spears
80 bretny spears
80 britnany spears
73 brinteny spears
73 brittainy spears
73 pritney spears
66 brintany spears
66 britnery spears
59 briitney spears
59 britinay spears

54 britneay spears
54 britner spears
54 britney's spears
54 britnye spears
54 britt spears
54 brttany spears
48 bitany spears
48 briny spears
48 brirney spears
48 britant spears
48 britnety spears
48 brittanny spears
48 brttney spears
44 birttany spears
44 brittani spears
44 brityney spears

Fuente: Google

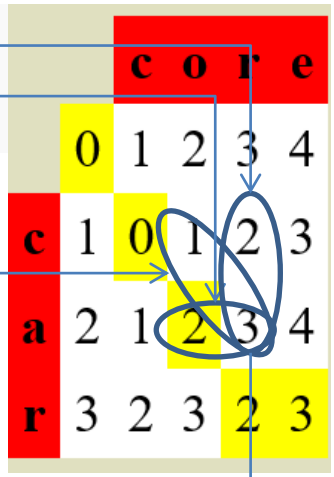
Distancia de Levehenstein

Corrección ortográfica comparando la diferencia entre dos cadenas, indica el esfuerzo en transformar una cadena en otra

Error	Correction	Correct letter	Error letter	Type
*acress	acres	-	s	insertion
*acress	actress	t	-	deletion
*acress	across	o	e	substitution
*acress	caress	ca	ac	transposition

```
Mínimo valor de:
d[i-1, j] + 1, // supresión
d[i, j-1] + 1, // inserción
d[i-1, j-1] + coste//sustitución
```

```
Si x(i)<>Y(j)→+2
Si f x(i)=Y(j)→0
```



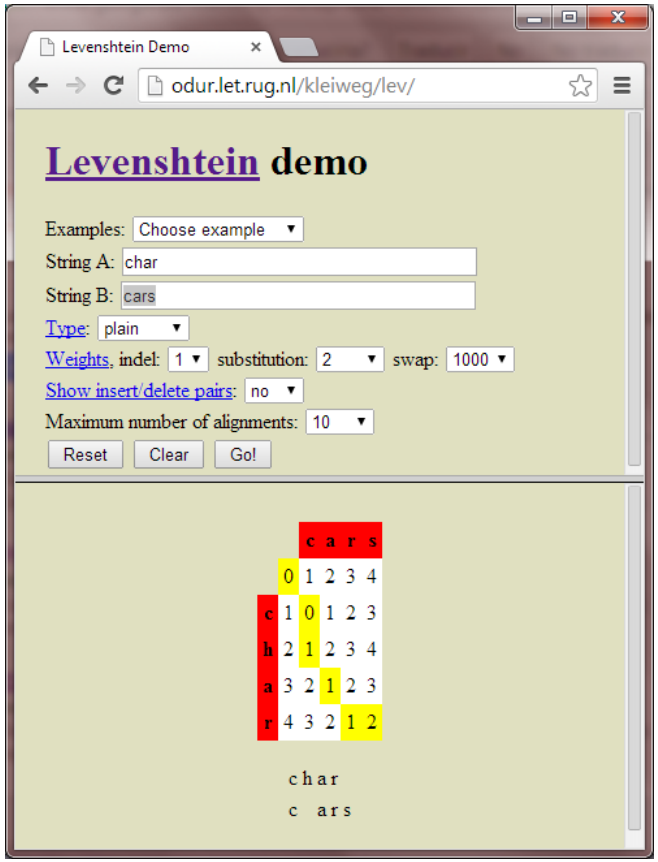
Distancia de edición d[3,4]

d [2,4] (r<>a)→2+1=3

d [3,3] (a<>r)→2+1=3

d [2,3] (r<>a) →1+2=3

Min=3



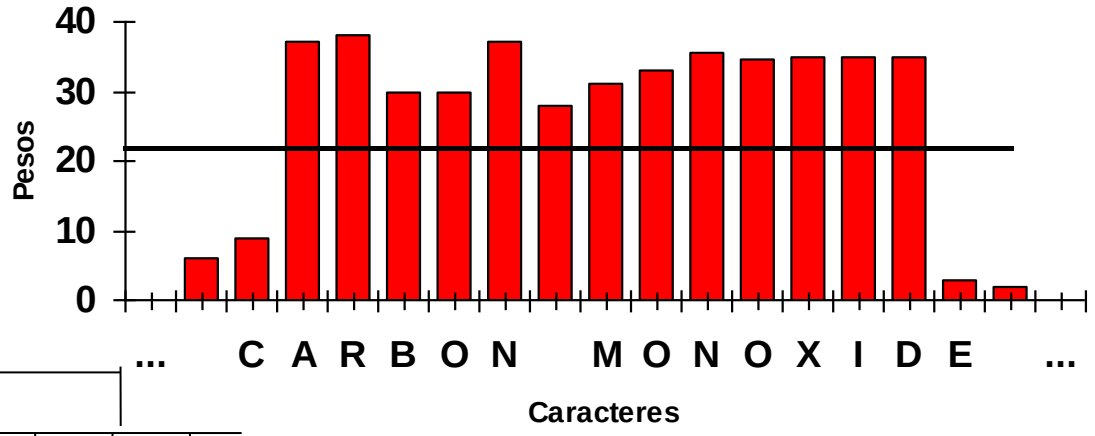
http://odur.let.rug.nl/kleiweg/lev/

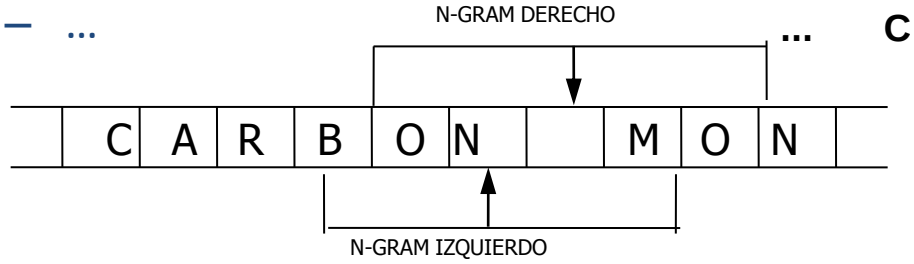
N-Grams. Frecuencia de subcadenas

- Probabilidad de que un conjunto de caracteres o palabras se produzcan. Si es baja puede ser un error tipográfico.
- Simple y escalable
- n-gram se puede aplicar a secuencia de caracteres o palabras (bigram si se calcula la frecuencia con la letra/palabra anterior, trigrama si dos letras/palabras anteriores, …)
- A veces con fingerprint (resultado de ordenar los caracteres alfabeticamente tras quitar duplicados)

Aplicaciones

- DNA
- Ortografia
- Alineamiento de datasets
- …





N-gram	http://guidetodatamining.com/ngramAnalyzer/	https://books.google.com/ngrams

Algoritmos fonéticos

Agrupan palabras que suenan parecido

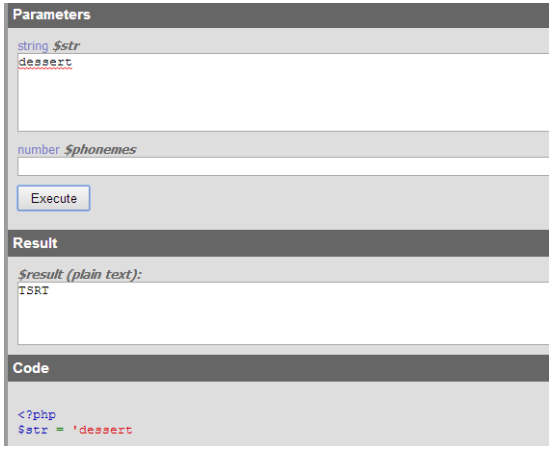
Algoritmos fonéticos: Soundex, Metaphone

- **Metaphone**

Sustituye conjuntos de letras por una concreta
http://www.php.net/manual/en/function.metaphone.php

Computer -> KMPTR

Prueba http://php.fnlist.com/string/metaphone:
seas, sees, seize / right, rite, wright, write / duel, jewel



- **Soundex**

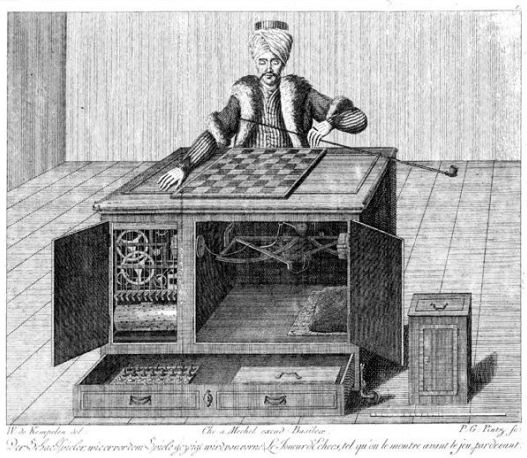
[http://www.gedpage.com/soundex.html]

[http://scriptun.com/php/online/soundex]

Computer-> C513

Crowdsourcing para limpieza y normalización

- Utilizar humanos como procesadores de un sistema distribuido
- Plataformas: Mturk (Mechanical Turk), Crowdflower, CloudCrowd, …
- Se definen tareas sencillas para realizar en Web (llmadas HITS:Human Intelligence Tasks) .
- Actualmente 200,000 workers
- Se hacen micropagos por HIT (5-20 c/HIT)
- Rápido y económico
- Hay formas de implementar control de calidad (preguntas trampa o conocidas, acuerdo entre anotadores, test de cualificación, ratio de aceptación, evitar superworkers ….) Similar a las folksonomías pero de pago



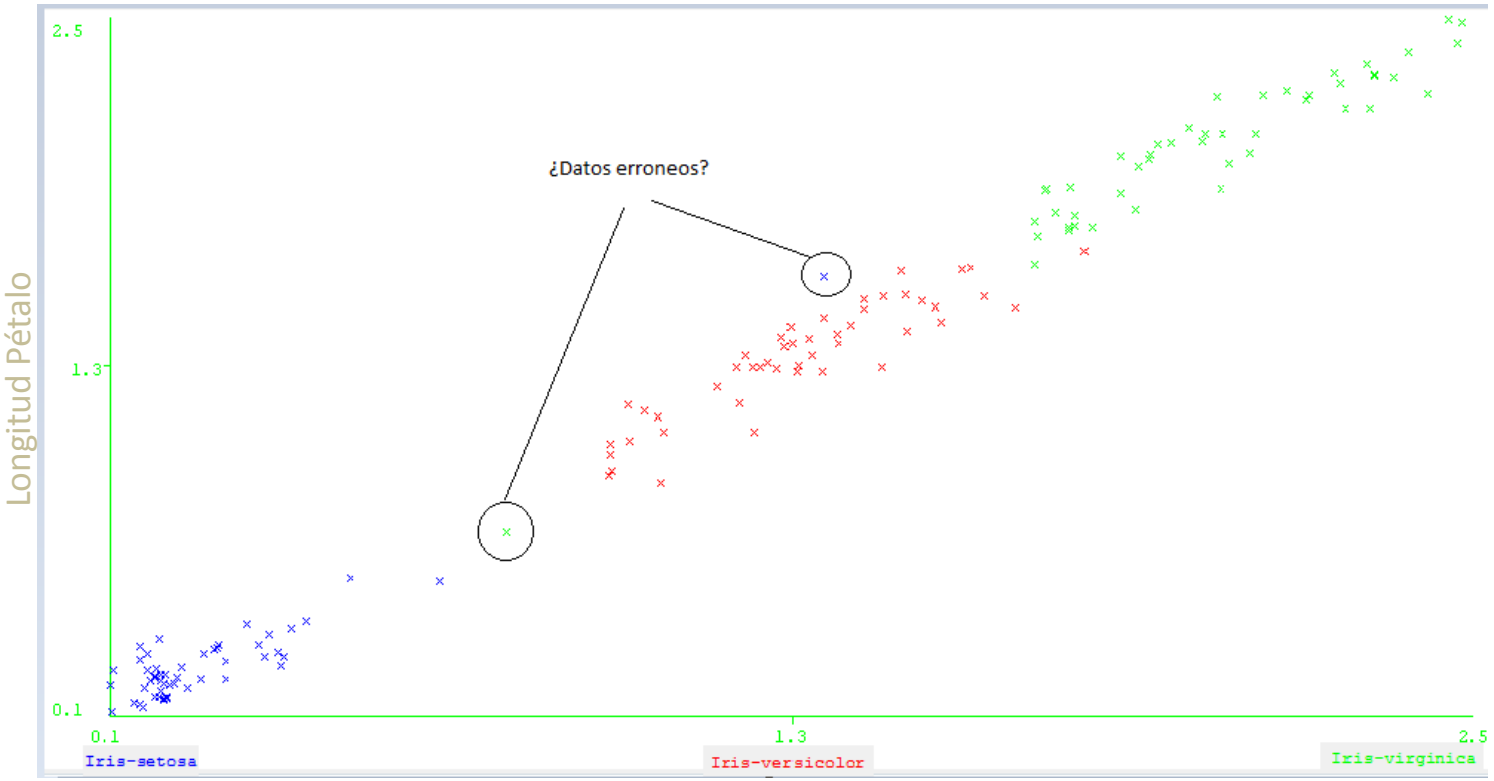
Proyectos para cleansing:
CrowdCleaning (U.Hong Kong)
Argonaut (U.Berkeley)
Katara (U.Waterloo)

A simple method to find duplicate records is to ask the crowd to check all possible pairs and decide whether each item in the pair refers to the same entity or not (CrowdER project)

Omar Alonso. Crowdsourcing 101

Problemas

Eliminación de información útil: modificar o eliminar datos puede suponer una perdida de información valiosa o generar nuevos errores. Sobre todo en procesos masivos automatizados.



Herramientas de limpieza y enriquecimiento

Muchas más, como Oracle, IBM (InfoSphere), Lavastorm, Experian, QAS Clean Web Service …

Herramientas	Funcionalidad
Google Refine (Open refine)	-Gran comunidad de usuarios - Fácil de instalar y de utilizar. - Levehstein, ngrams, fingerprint, …. -- Número creciente de mejoras -- Capacidad de incrementar la información con datos relacionados (Reconciliation) -- Corrección errores comunes en datasets creados manualmente -- Extensiones para crowsourcing y NER
DataWrangler	Proyecto finalizado Sugiere acciones según anteriores procesos Fácil de usar aunque poca documentación
Pentaho- Kettle	Realiza ETL. Pentaho Data Integration Se pueden crear reglas para depurar, extraer y normalizar información
QLink	Herramienta de visualización de patrones utilizada en el sector financiero

Links

- Analysing UK Lobbying Data Using OpenRefine http://schoolofdata.org/2013/06/04/analysing-uk-lobbying-data-using-openrefine/
- Harvesting and Analyzing Tweets http://schoolofdata.org/harvesting-and-analyzing-tweets/
- From Excel file to RDF with links to DBpedia and Europeana (from around the web) http://googlerefine.blogspot.com.es/2012/11/from-excel-file-to-rdf-with-links-to.html
- Google Refine tutorial http://davidhuynh.net/spaces/nicar2011/tutorial.pdf
- Questioning Election Data to see if it has a story to tell http://blog.ouseful.info/2013/05/05/questioning-election-data-to-see-if-it-has-a-story-to-tell/
- Resources: Data journalism toolbox https://fellows.knightscience.org/tutorials/resources-data-journalism-toolbox/
- Big Data – What it means for the digital analyst http://online-behavior.com/analytics/big-data

Contenido

- Ciclo de Vida de la Información
- Fuentes de la Información
- Estructuración y Saneamiento de los datos: la coherencia
- Integración de fuentes

Integración de Fuentes. Data Mapping

- Se pueden fusionar campos de distintas fuentes mediante elementos comunes.
- Estos elementos comunes pueden serlo por:
  - Elementos relacionados semánticamente: por ejemplo si dos fuentes tienen un campo indicando el país, se podría relacionar por ese campo.
  - También hay aplicaciones y algoritmos para fusionar tablas. Se analizan los términos y se ven conjuntos de términos similares en otros recursos candidatos.
  - Recursos que manualmente tienen correspondencia. Por ejemplo, el proyecto Linked Data tiene miles de recursos que han sido manualmente enlazados

Interconexión de datos

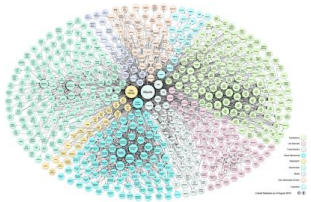
	Estrategia de adquisición de datos	Problema	Contexto
Mediados 90	Herramientas ETL para normalizar y limpiar DCOM or CORBA para intercambio de datos	Muchos datos sucios Sin protocolos, soluciones contradictorias Bases de datos no conectados, aplicaciones aisladas	SQL
Año 2000	SOAP framework	Muy complejo Poca compatibilidad con JS Esquemas en XML	Internet XML and XQuery RDF
Mediados 2000	MapReduce NoSQL		JSON Hadoop

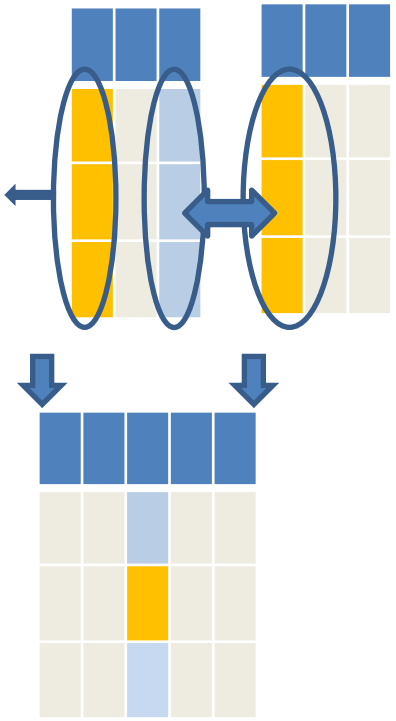
JSON, fácil de leer y compacto

Enriquecimiento y conexión de datos

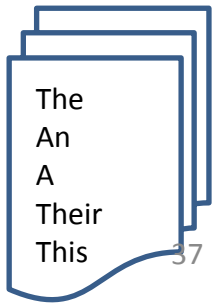
Estrategias para emparejar un dataset y un dato del LoD.

- Estrategias para Integrar y enriquecer :
  - [además de los de semejanza léxica]
  - Eliminar palabras vacías
  - Comparar con minúsculas y sin acentos
  - Stemming o lematización
  - Comprobar contra listados, vocabularios u ontologías
- Agrupamiento
  - Listas de sinónimos
  - Anotaciones
  - Distancias semánticas en tesauros y ontologías
  - Algoritmos de clustering





Palabras vacías



Módulo IV
Acceso y recuperación de datos en la Web

Colaboradores

J.Morato, V.Palacios