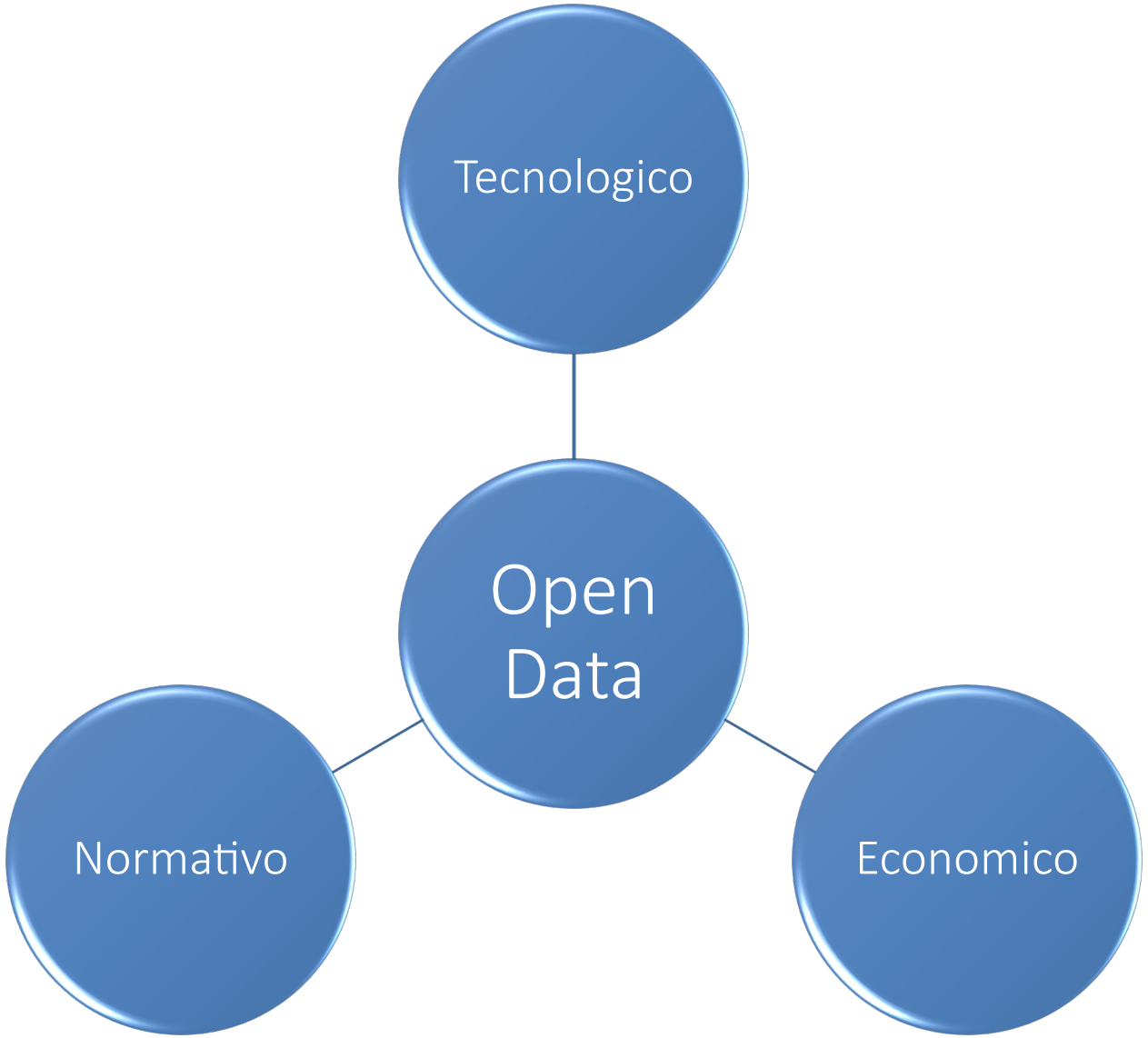
Open Data Management

La classificazione dei dati aperti

Ing. Davide Taibi

I dati aperti… non solo una questione tecnologica



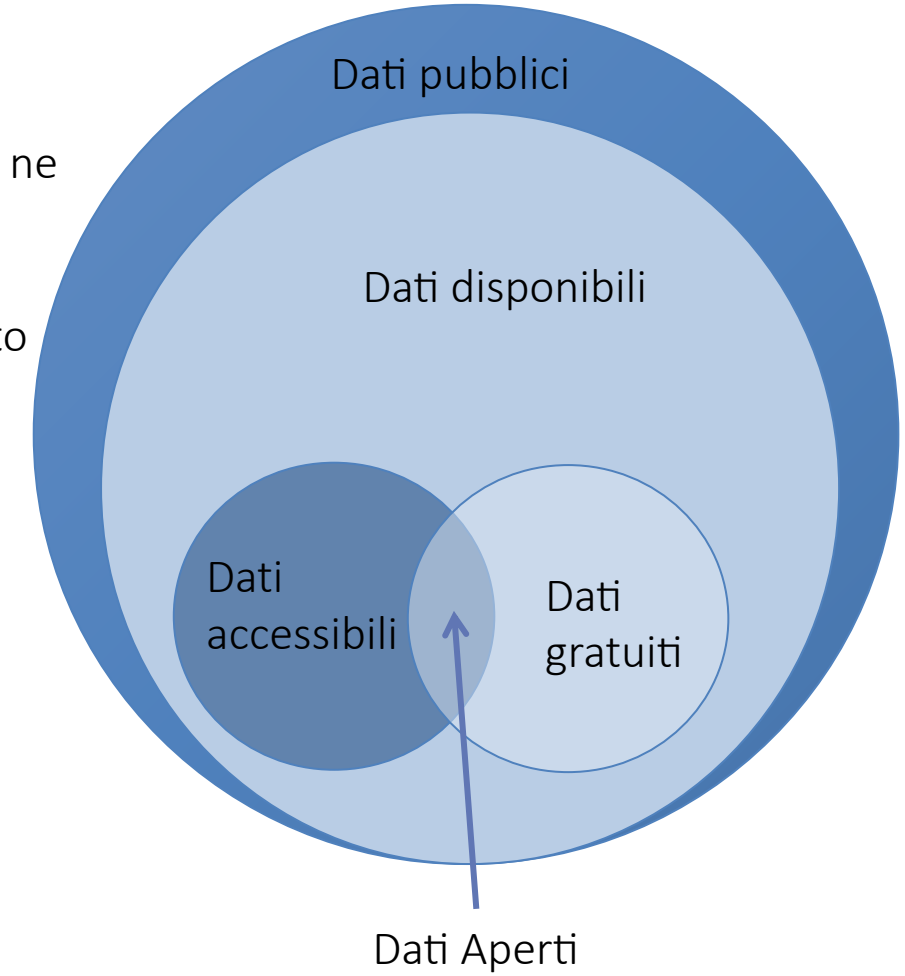
Tipi di dato e Dati Aperti

Linee Guida recanti regole tecniche per l'apertura dei dati e il riutilizzo dell'informazione del settore pubblico (2023)

Linee Guida per la Valorizzazione del Patrimonio Informativo Pubblico (2016)

Dato aperto (risponde a tre requisiti):

- **Disponibile** (requisito giuridico) secondo I termini di una licenza che ne permetta l'utilizzo da parte di chiunque, anche per finalità commerciali, in formato disaggregato
- **Accessibile** (requisito tecnologico) attraverso le tecnologie dell'informazione e della comunicazione, in formato aperto e con i relativi metadati
- **Gratuito** (requisito economico): disponibili gratuitamente oppure …



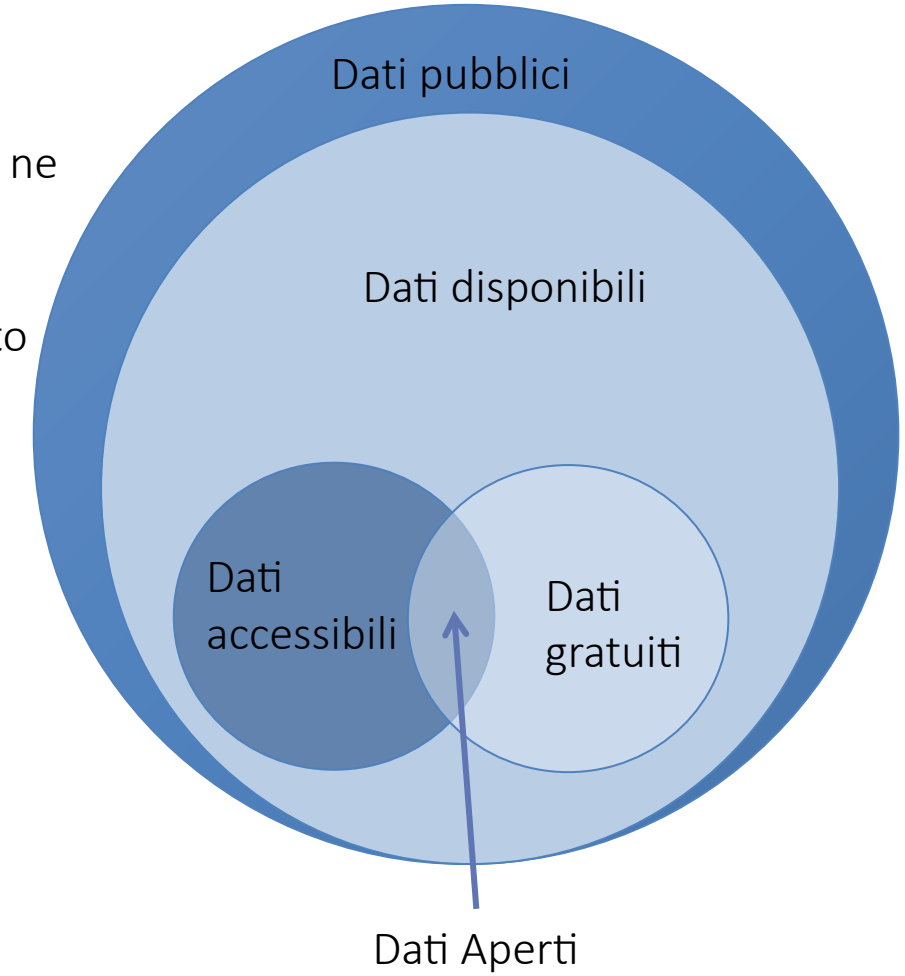
Tipi di dato e Dati Aperti

Linee Guida recanti regole tecniche per l'apertura dei dati e il riutilizzo dell'informazione del settore pubblico (2023)

Linee Guida per la Valorizzazione del Patrimonio Informativo Pubblico (2016)

Dato aperto (risponde a tre requisiti):

- **Disponibile** (requisito giuridico) secondo I termini di una licenza che ne permetta l'utilizzo da parte di chiunque, anche per finalità commerciali, in formato disaggregato
- **Accessibile** (requisito tecnologico) attraverso le tecnologie dell'informazione e della comunicazione, in formato aperto e con i relativi metadati
- **Gratuito** (requisito economico): disponibili gratuitamente oppure …



disponibili ai costi marginali sostenuti per la loro riproduzione, messa a disposizione e divulgazione. AgID, su proposta dell'amministrazione titolare, determina le tariffe standard e le pubblica sul proprio sito istituzionale.
Eccezione: dati per i quali le pubbliche amministrazioni e gli organismi di diritto pubblico generano utili sufficienti per coprire una parte sostanziale dei costi di raccolta, produzione, riproduzione e diffusione. Con decreti dei Ministeri competenti, di concerto con il Ministero dell'economia e delle finanze, sentita AgID, si determinano le tariffe e le modalità di versamento a fronte delle suddette attività.

Dati Aperti

L'art. 68, c. 3, del d.lgs. n. 82/2005 (come sostituito dall'art. 9, c. 1, lett. b), d.l. n. 179/2012, convertito con modificazioni, dall'art. 1, c. 1, l. n. 221/2012), e aggiornato dal d.lgs 217/2017 entrato in vigore il 27 gennaio 2018, definisce come dati di tipo aperto quelli che presentano le seguenti caratteristiche:

1) sono disponibili secondo i termini di una licenza o di una previsione normativa che ne permetta l'utilizzo da parte di chiunque, anche per finalità commerciali, in formato disaggregato;

2) sono accessibili attraverso le tecnologie dell'informazione e della comunicazione, ivi comprese le reti telematiche pubbliche e private, in formati aperti, sono adatti all'utilizzo automatico da parte di programmi per elaboratori e sono provvisti dei relativi metadati;

3) sono resi disponibili gratuitamente attraverso le tecnologie dell'informazione e della comunicazione, ivi comprese le reti telematiche pubbliche e private, oppure sono resi disponibili ai costi marginali sostenuti per la loro riproduzione e divulgazione.

Serie di dati di elevato valore

La Direttiva (UE) 2019/1024 del Parlamento europeo e del Consiglio, del 20 giugno 2019, relativa all'apertura dei dati e al riutilizzo dell'informazione del settore pubblico e, quindi, il Decreto Legislativo 8 novembre 2021, n. 200 relativa alla sua attuazione hanno introdotto il concetto di «**serie di dati di elevato valore**»

Sono quei: "*documenti il cui riutilizzo è associato a importanti benefici per la società, l'ambiente e l'economia, in considerazione della loro idoneità per la creazione di servizi, applicazioni a valore aggiunto e nuovi posti di lavoro, nonché del numero dei potenziali beneficiari dei servizi e delle applicazioni a valore aggiunto basati su tali serie di dati*".

Serie di dati di elevato valore

I criteri per l'identificazione di tale tipologia di dati sono indicati all'art. 14, comma 2 della Direttiva, secondo cui deve essere valutata la loro potenzialità:

a) nell'apportare importanti benefici socio-economici o ambientali e servizi innovativi;

b) nel beneficiare un numero elevato di utilizzatori, in particolare PMI;

c) nel contribuire a generare proventi;

d) nell'essere combinati con altre serie di dati.

Serie di dati di elevato valore

Le categorie tematiche delle serie di dati di elevato valore indicate nell'Allegato I della Direttiva sono:

1. Dati geospaziali
2. Dati relativi all'osservazione della terra e all'ambiente
3. Dati meteorologici
4. Dati statistici
5. Dati relativi alle imprese e alla proprietà delle imprese
6. Dati relativi alla mobilità.

Dati dinamici

dati dinamici: documenti informatici, soggetti ad aggiornamenti frequenti o in tempo reale, in particolare a causa della loro volatilità o rapida obsolescenza (art 2 c.1 l c-exes)

Es. i dati generati da sensori sono solitamente considerati dati dinamici (art. 2, p. 8) per cui esempi di dati dinamici includono: i dati ambientali, relativi al traffico, satellitari o meteorologici.

I dati dinamici DEVONO essere resi disponibili per il riutilizzo immediatamente dopo la raccolta tramite API adeguate e, ove possibile, attraverso download in blocco.

Dati della ricerca

I dati della ricerca sono definiti dal d.l. 8 novembre 2021, n. 200 come

“documenti informatici, diversi dalle pubblicazioni scientifiche, raccolti o prodotti nel corso della ricerca scientifica e utilizzati come elementi di prova nel processo di ricerca, o comunemente accettati nella comunità di ricerca come necessari per convalidare le conclusioni e i risultati della ricerca”.

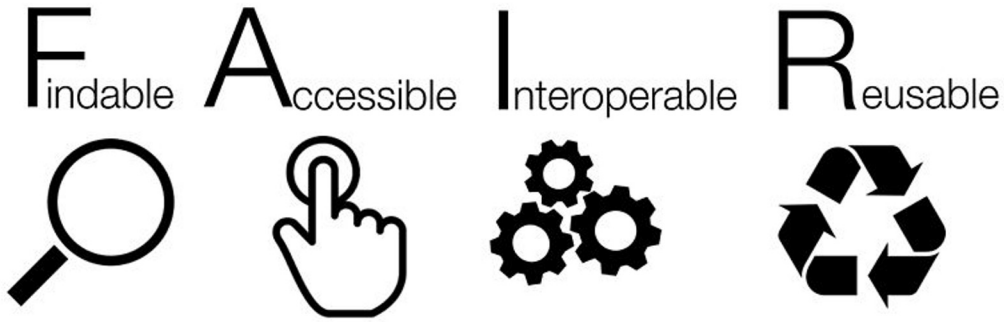
Es: statistiche, risultati di esperimenti, misurazioni, osservazioni risultanti dall’indagine sul campo, risultati di indagini, immagini e registrazioni di interviste, oltre a metadati, specifiche e altri oggetti digitali

Dati della ricerca

Per identificare i dati della ricerca da rendere disponibili per il riutilizzo, è necessario tenere conto della normativa in materia di protezione dei dati personali, degli interessi commerciali, dei diritti di proprietà intellettuale e dei diritti di proprietà industriale.

La Direttiva UE indica che bisogna considerare il principio 'as open as possible, as closed as necessary' -> "*il più aperto possibile, chiuso il tanto necessario*"

I dati della ricerca DEVONO rispettare i 4 principi del framework **FAIR** (Findable- Accessible- Interoperable- Reusable) attraverso i requisiti di reperibilità, accessibilità, interoperabilità e riutilizzabilità



«…le macchine devono essere in grado di agire in modo autonomo e appropriato in relazione alla vasta gamma di tipi, formati e meccanismi di accesso/protocolli che incontreranno quando esplorano l'ecosistema di dati globale»

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Findable (Reperibile)- Rendere i dati reperibili da macchine ed essere umani. Per fare questo, dovrebbero essere resi disponibili i metadati attraverso una risorsa consultabile online e dovrebbe essere assegnato un identificatore persistente a dati e metadati.

Accessible (Accessibile)- Deve essere possibile ad essere umani e macchine accedere ai dati attraverso protocolli standard e aperti.

Interoperable (Interoperabile)- Dati e metadati devono poter essere combinati con altri dati e/o strumenti. Per questo, devono essere conformi a formati e standard riconosciuti

Reusable (Riutilizzabile)- I dati devono essere ben documentati in modo che possano essere interpretati correttamente, replicati e/o combinati anche in contesti diversi. Ai dati, inoltre, bisogna assegnare una licenza chiara e accessibile in modo che si possa capire che tipo di riutilizzo è consentito. Resta fermo il dovuto rispetto dei limiti al riutilizzo derivanti dalla normativa europea e nazionale in materia di protezione dei dati personali.

Dati territoriali

I dati territoriali sono definiti dal CAD come "i dati che attengono, direttamente o indirettamente, a una località o a un'area geografica specifica"

Direttiva INSPIRE

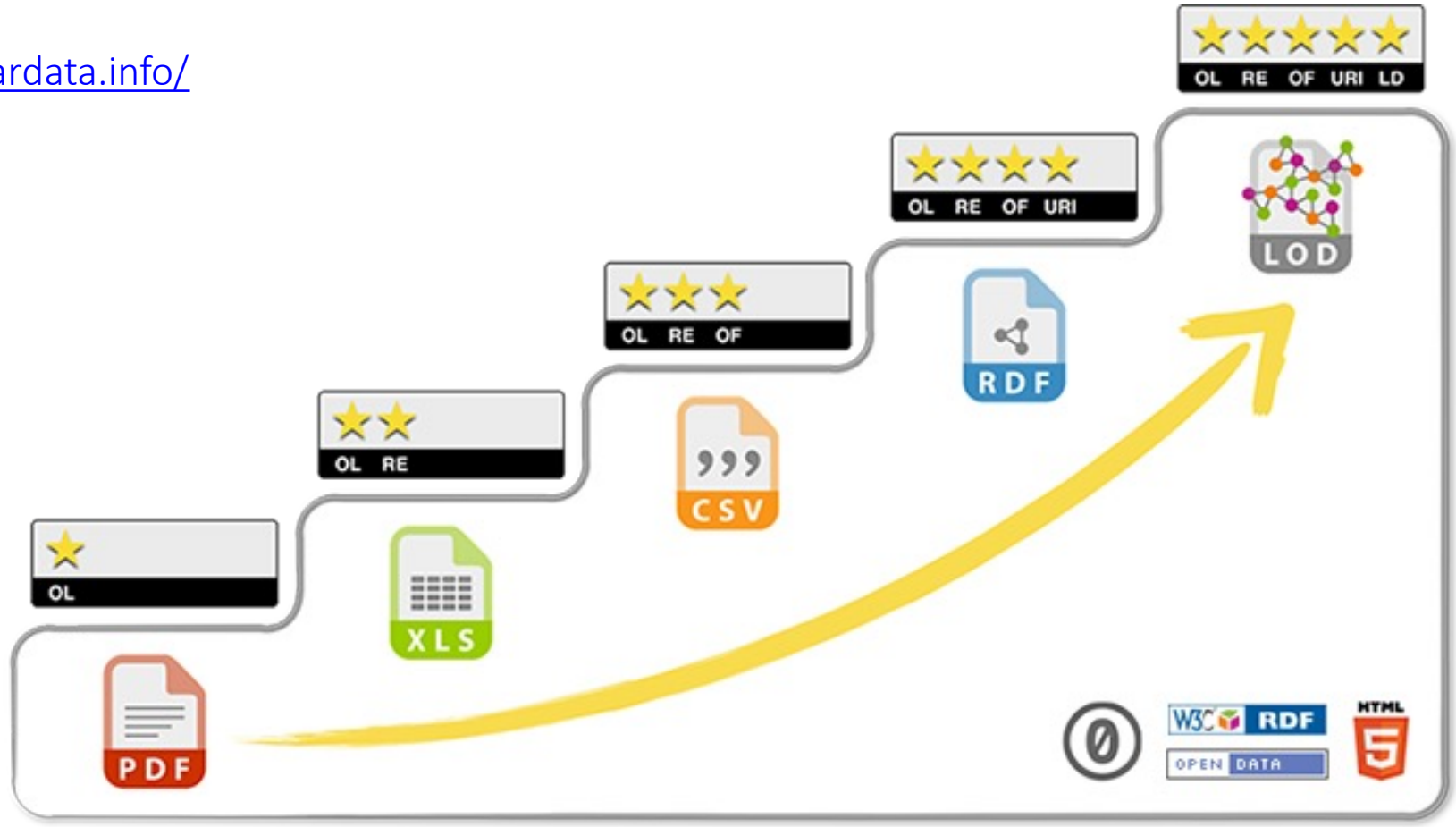
Dati di elevato valore: "Dati geospaziali", "Dati relativi all'osservazione della terra e all'ambiente", "Dati meteorologici", "Dati relativi alla mobilità"

I dati territoriali resi disponibili per il riutilizzo DEVONO essere documentati esclusivamente attraverso metadati conformi alle *"Linee Guida recanti regole tecniche per la definizione e l'aggiornamento del contenuto del Repertorio Nazionale dei Dati Territoriali"* e le relative guide operative.

https://geodati.gov.it/geoportale/

I 5 livelli

http://5stardata.info/



I 5 livelli

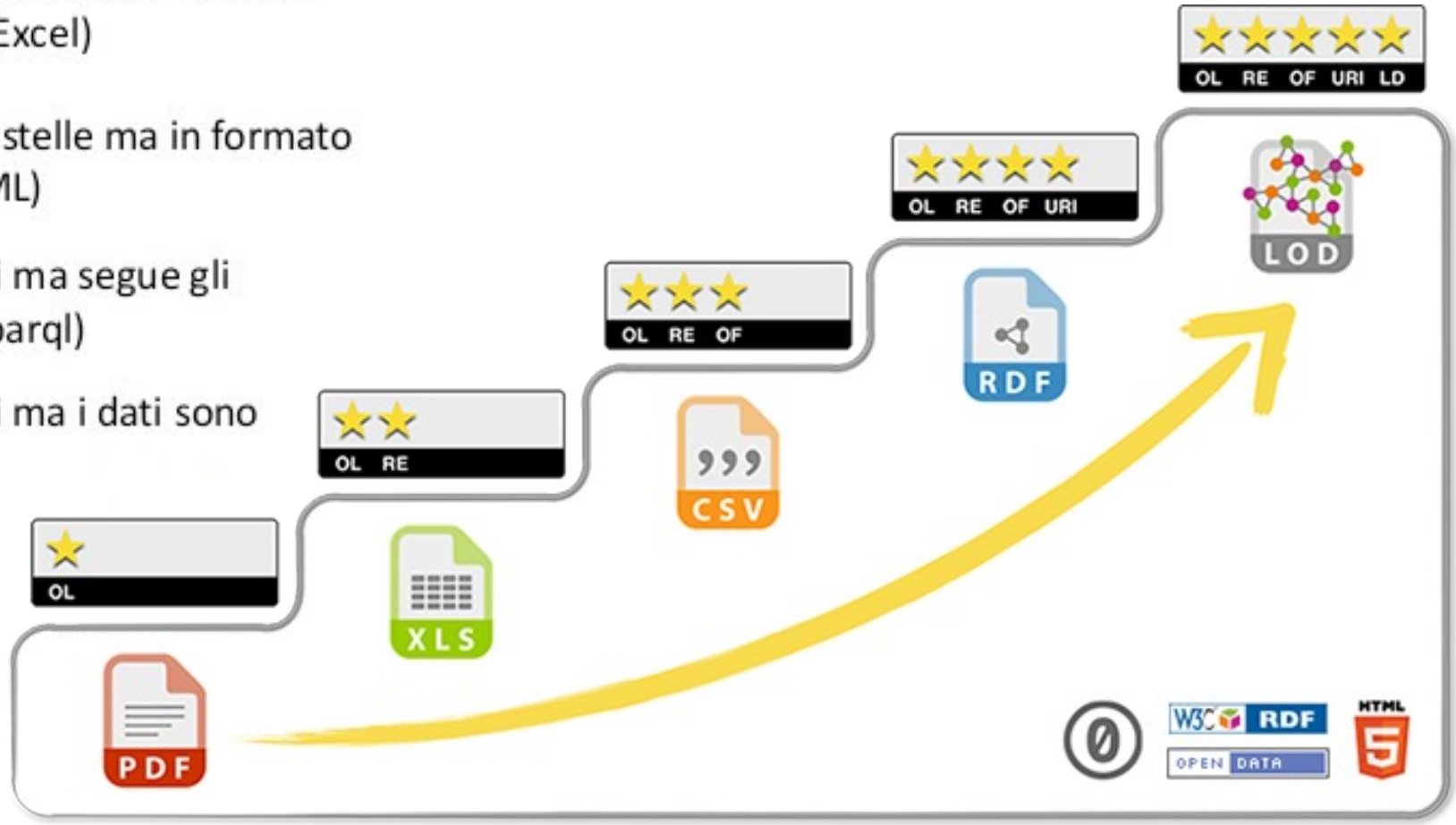
Dato disponibile sul Web in un qualsiasi formato (anche PDF) rilasciato con licenza Open

Leggibile dal calcolatore. Dati strutturati in formati proprietari (es. Excel)

Come gli Open Data a 2 stelle ma in formato non proprietario (es. XML)

Come I livelli precedenti ma segue gli standard W3C (RDF e Sparql)

Come i livelli precedenti ma i dati sono collegati (Linked Data)



I 5 livelli

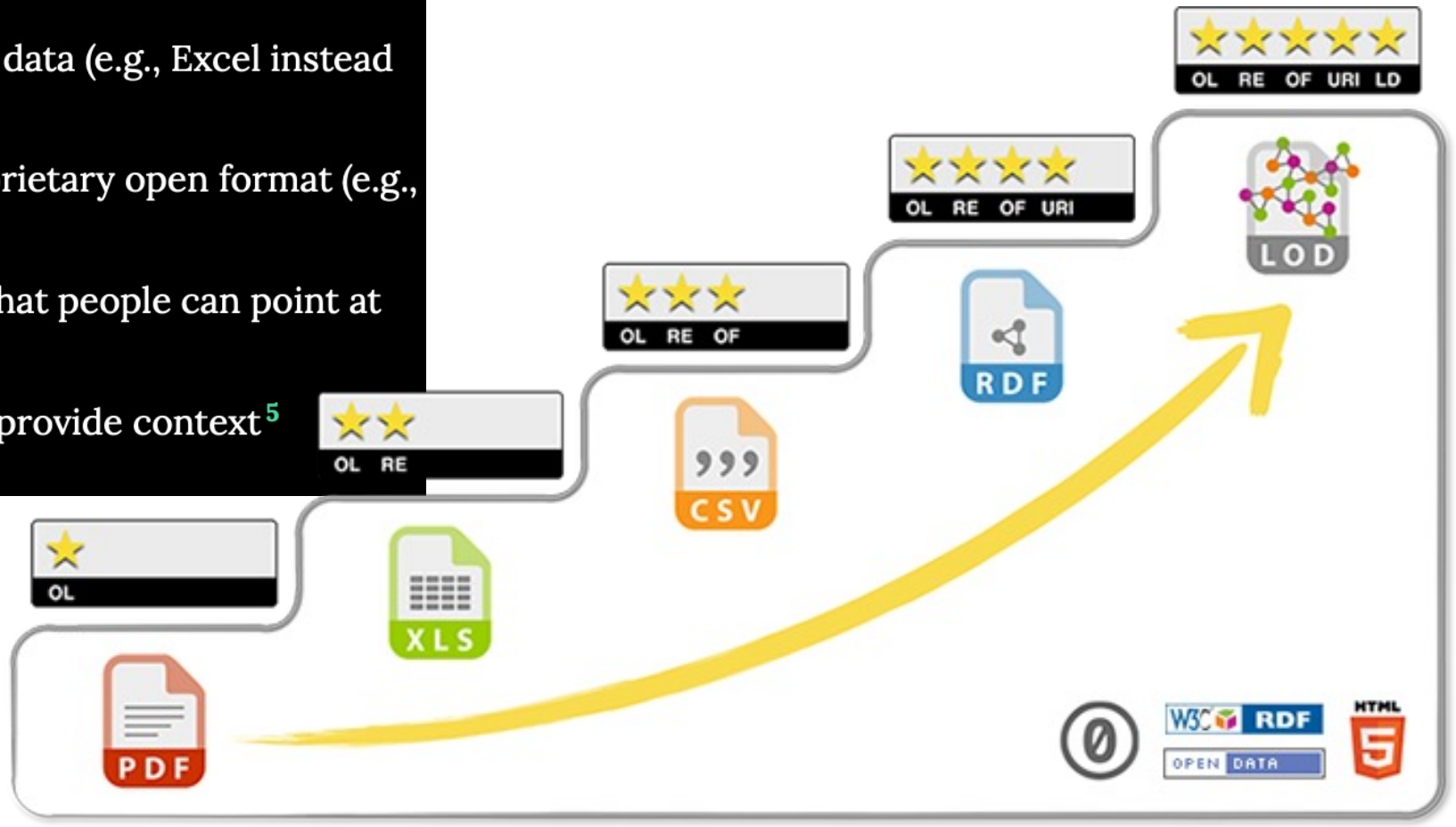
★ make your stuff available on the Web (whatever format) under an open license[1]

★★ make it available as structured data (e.g., Excel instead of image scan of a table)[2]

★★★ make it available in a non-proprietary open format (e.g., CSV instead of Excel)[3]

★★★★ use URIs to denote things, so that people can point at your stuff[4]

★★★★★ link your data to other data to provide context[5]



URI, URL, URN

URI: Universal Resource Identifier

Un Uniform Resource Identifier (URI) è una sequenza di caratteri che identifica una risorsa astratta o fisica.

URL: Universal Resource Locator

Il termine "Uniform Resource Locator" (URL) si riferisce al sottoinsieme di URI che, oltre a individuare una risorsa, forniscono un mezzo per localizzare la risorsa descrivendo il meccanismo di accesso principale.

URN: Universal Resource Name

Il termine "Uniform Resource Name" (URN) è stato utilizzato storicamente per riferirsi sia alle URI nello schema "urn" [RFC2141], tenute a rimanere uniche a livello globale e persistenti anche quando la risorsa cessa di esistere o non è più disponibile. Un URN può quindi essere usato per identificare una risorsa, senza lasciarne intendere l'ubicazione o come ottenerne una rappresentazione. Per esempio l'URN urn:isbn:0-395-36341-1 è un URI che mappa universalmente e univocamente un libro mediante il suo identificativo, o nome, (0-395-36341-1) nel namespace dei codici ISBN, ma non suggerisce dove e come possiamo ottenere una copia di tale libro.

Fonte: RFC 3986 **Uniform Resource Identifier (URI): Generic Syntax**

Esempi di URI

ftp://ftp.is.co.za/rfc/rfc1808.txt -- schema per servizi FTP

http://www.math.uio.no/faq/compression-faq/part1.html -- schema per servizi HTTP

mailto:mduerst@ifi.unizh.ch -- schema per indirizzi di posta elettronica

news:comp.infosystems.www.servers.unix -- schema per newsgroup e articoli Usenet

telnet://melvyl.ucop.edu/ -- schema per servizi interattivi telnet

irc://irc.freenode.net/wikipedia-it -- schema per IRC

Fonte Wikipedia- https://it.wikipedia.org/wiki/Uniform_Resource_Identifier

URL- schema

URL – RFC 1738

scheme:[//[user:password@]host[:port]][/]path[?query][#fragment]

URI persistenti

Fonte: Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC

URI persistenti

Da evitare URI contenenti:

- nome del progetto/ufficio/unità amministrativa che detiene la risorsa per evitare problemi derivanti dalla fine del progetto stesso o fusioni o chiusure di uffici nell'organizzazione;

- numeri di versione;

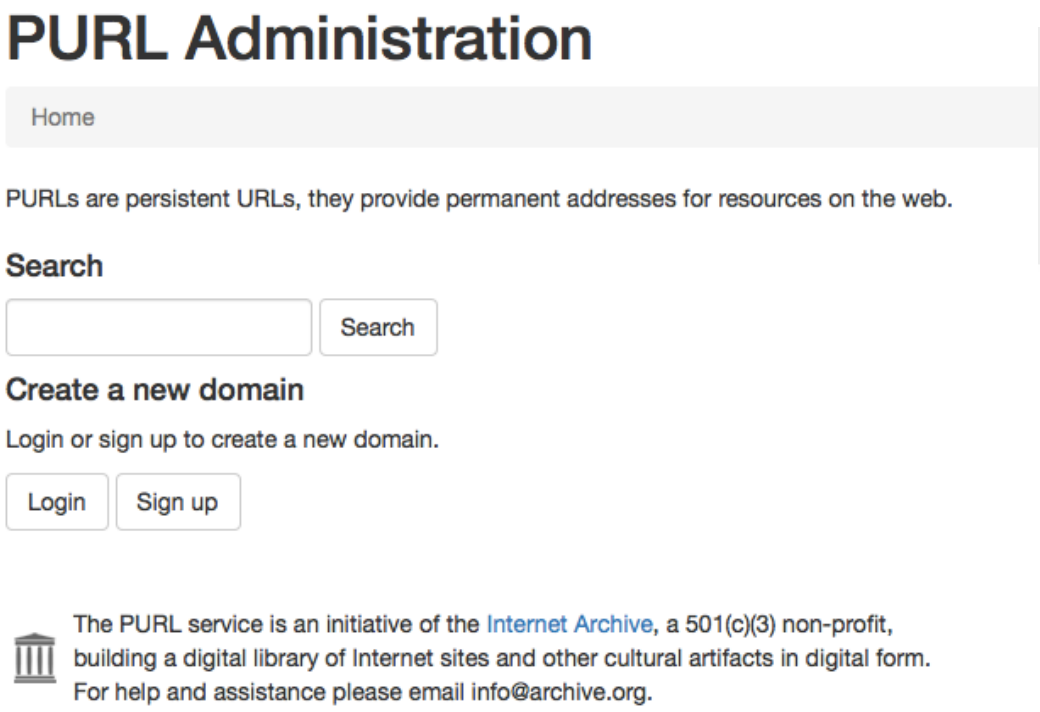
- identificatori esistenti che in passato sono stati utilizzati per identificare risorse differenti;

- Riferimenti generati in modo automatico e incrementale a meno che non vi sia la garanzia che il processo non venga mai più ripetuto o, se ripetuto, generi sicuramente gli stessi identificatori per gli stessi dati di input;

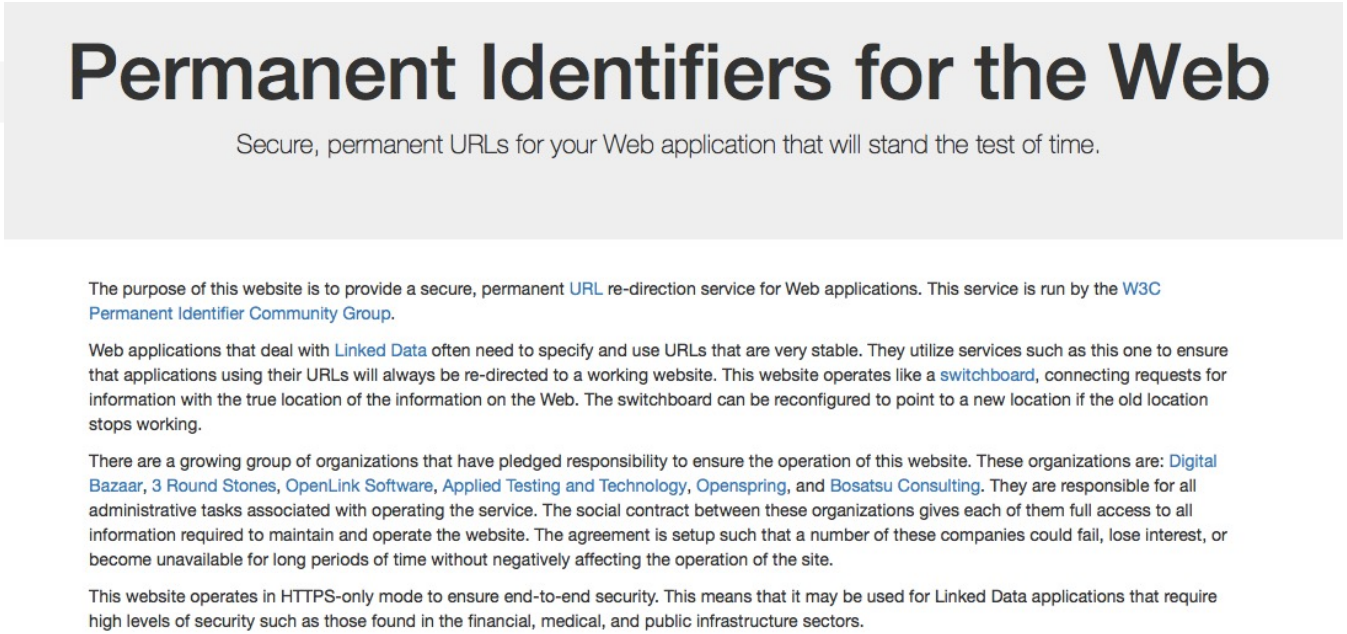
- stringhe rappresentanti "query" a database;

- Estensione del file.

URI persistenti

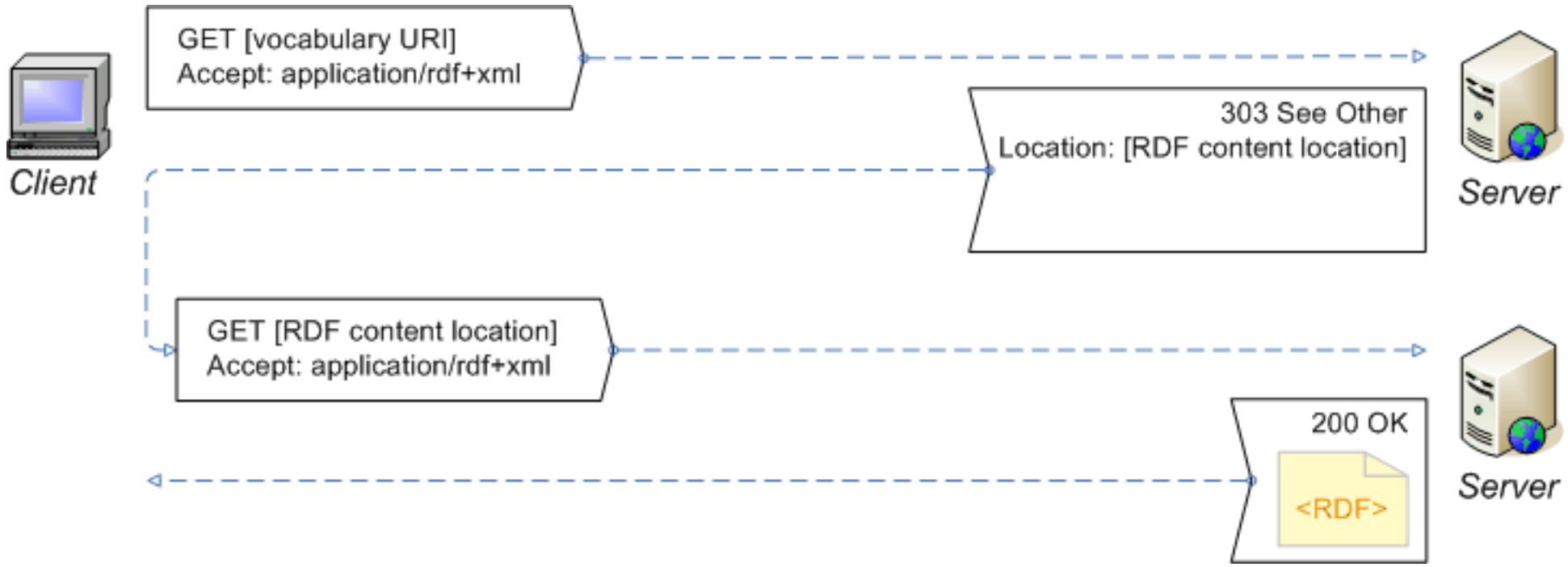


purl.org



w3id.org

Dereferencing HTTP URIs

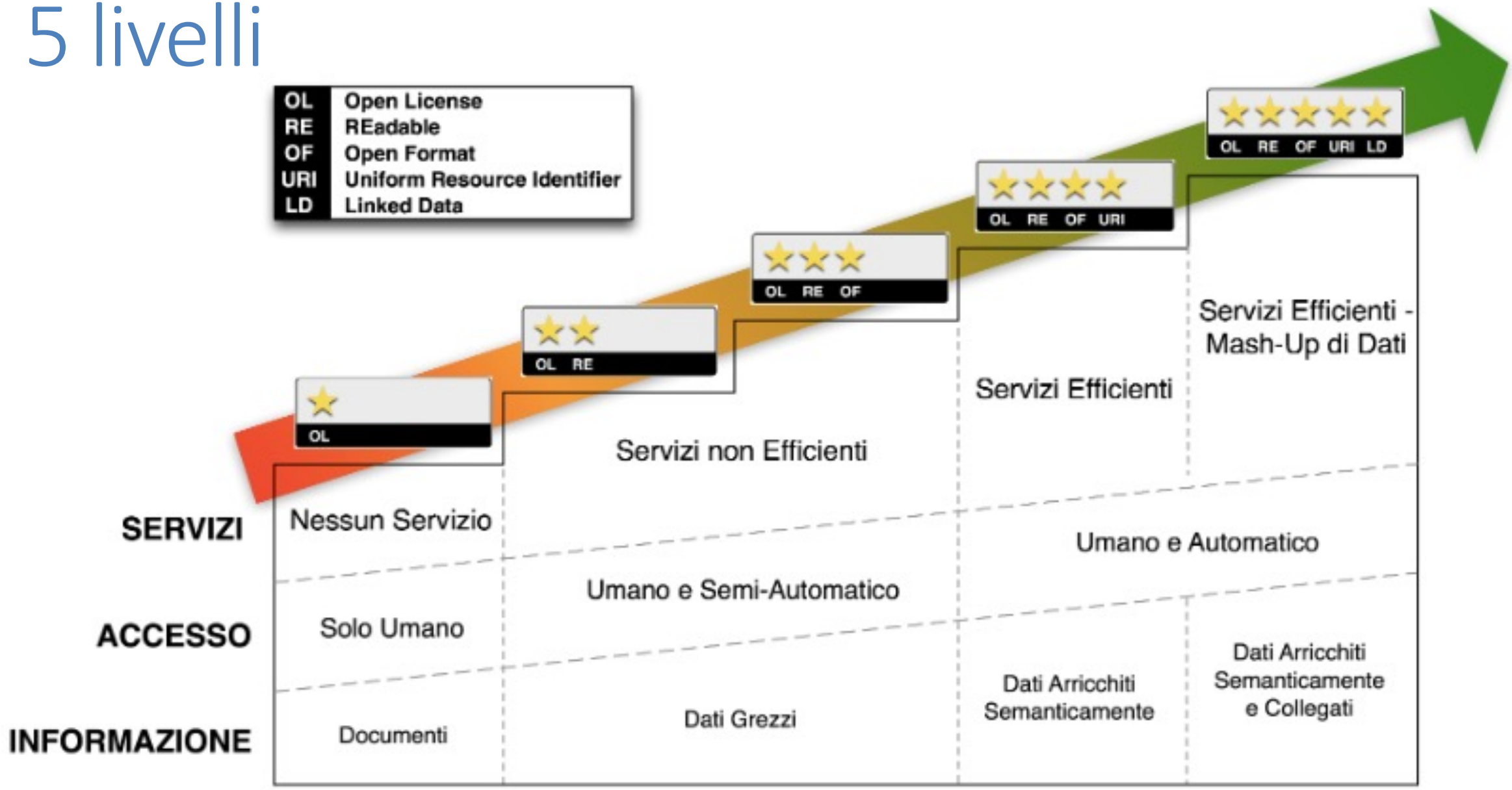


Open Data Management

La classificazione dei dati aperti, formati e metadati

Ing. Davide Taibi

I 5 livelli



Fonte: Linee Guida per la Valorizzazione del Patrimonio Informativo Pubblico (2016)

Formati per i documenti

Formati OASIS (Open Document Format for Office Applications)**:**

- **ODT (Open Document Text)**
- **ODS (Open Document Spreadsheet)**
- **ODP (Open Document Presentation)**

Formati principali in OpenOffice.org e LibreOffice.
Supportati da Microsoft Office, Google Drive e IBM Lotus.

https://www.oasis-open.org/

Formati per i documenti

PDF (Portable Document Format). Creato da Adobe, standardizzato dall'ISO (ISO/IEC 32000-1:2008). Diverse specifiche:

- PDF/A (PDF/Archiving) per l'archiviazione a lungo termine;
- PDF/X (PDF/eXchange) per le arti grafiche e la prestampa;
- PDF/E (PDF/Engineering) per la documentazione di tipo ingegneristico;
- PDF/H (PDF/Healthcare) per il settore sanitario;
- PDF/UA (PDF/Universal Accessibility) per l'accessibilità.

Akoma Ntoso. Linguaggio XML per la rappresentazione di documenti giuridici. È utilizzato dal Parlamento Europeo e dalla Commissione Europea come standard documentale per i documenti legislativi, giuridici e allegati tecnici.

http://www.akomantoso.org

Formati per i dati geografici

Shapefile: formato standard de-facto per la rappresentazione dei dati dei sistemi informativi geografici (GIS).

I dati sono di tipo vettoriale. Lo shapefile è stato creato dalla società privata ESRI che rende comunque pubbliche le sue specifiche.

L'apertura delle specifiche ha consentito lo sviluppo di diversi strumenti in grado di gestire e creare tale formato.

Impropriamente ci si riferisca a uno shapefile, nella pratica si devono considerare almeno tre file:

- un .shp contenente le forme geometriche,

- un .dbf contenente il database degli attributi delle forme geometriche

- un .shx come indice delle forme geometriche.

A questi tre si deve anche accompagnare un file .prj che contiene le impostazioni del sistema di riferimento.

Formati per i dati geografici

KML. È un formato basato su XML per rappresentare dati geografici.

Creato da Google, è diventato standard OGC (Open Geospatial Consortium).

Le specifiche della versione 2.2 presentano una serie di entità XML attraverso cui archiviare le coordinate geografiche che rappresentano punti, linee e poligoni espressi in coordinate WGS84 e altre utili a definire gli stili attraverso cui i dati andranno visualizzati.

Distribuito anche in modalità compressa attraverso file con estensione .kmz.

Formati per i dati geografici

GeoJSON

E’ un formato aperto per la rappresentazione e l'interscambio dei dati territoriali in forma vettoriale, basato su JSON (JavaScript Object Notation).

Ogni dato è codificato come oggetto, e a ogni oggetto è associato un insieme di coppie nome/valore (membri).

I principali nomi di membri che rappresentano le caratteristiche dei dati geografici sono:

- "**type**" che serve ad indicare il tipo di geometria (punto, linea, poligono o insieme multi-parte di questi tipi);
- "**coordinates**" attraverso cui sono indicate le coordinate dell'oggetto in un dato sistema di riferimento;
- "**bbox**" attraverso cui sono indicate le coordinate di un riquadro di delimitazione geografica;
- "**crs**" (opzionale) per l'indicazione del sistema di riferimento.

Formati per i dati geografici

GML (Geography Markup Language).

Basato su XML, rappresenta un formato di scambio aperto per i dati territoriali.

Definito originariamente da OGC e diventato lo Standard ISO 19136:2008, fornisce la codifica XML (schemi XSD) delle classi concettuali definite in diversi Standard ISO della serie 19100 e di classi aggiuntive appositamente definite: geometrie, oggetti topologici, unità di misura, tipi di base, riferimenti temporali, feature, sistemi di riferimento, copertura.

Formati per i dati geografici

GeoPackage

Formato aperto per la rappresentazione di dati geografici che può essere considerato un’alternativa al formato shapefile.

Esso supporta SpatiaLite ovvero una estensione dello schema del database SQLite.

Il principale vantaggio offerto da GeoPackage è quello di poter rappresentare in un unico file diversi dati geografici, sia di tipo vettoriali che raster, che possono essere gestiti anche tramite apposite query SQL.

Lo standard GeoPackage è riconosciuto dall'Open Geospatial Consortium (OGC).

Formati generici

XML

CSV (appendice b)

JSON / **JSON-LD**

Notation3 (N3) e Turtle (versione semplificata di N3)

N-Triples

Altri formati

- Dati relativi ai trasporti
- Dati Statistici
- Dati metereologici

General Transit Feed Specification (GTFS)

General Transit Feed Specification (**GTFS**) definisce un formato comune per gli orari dei trasporti pubblici e le relative informazioni geografiche.

Un feed GTFS è una collezione di file CSV contenuta in una file zip. Insieme le tabelle CSV correlate descrivono delle operazioni sulle tabelle degli orari del sistema di trasporto.

Le specifiche sono progettate per essere sufficienti a fornire funzionalità di pianificazione di un viaggio, ma è anche utile per altre applicazioni come analisi del livello di servizio e altre misure prestazionali. Non include informazioni in tempo reale sebbene possano essere correlate con le specifiche di GTFS-realtime.

(fonte: Wikipedia)

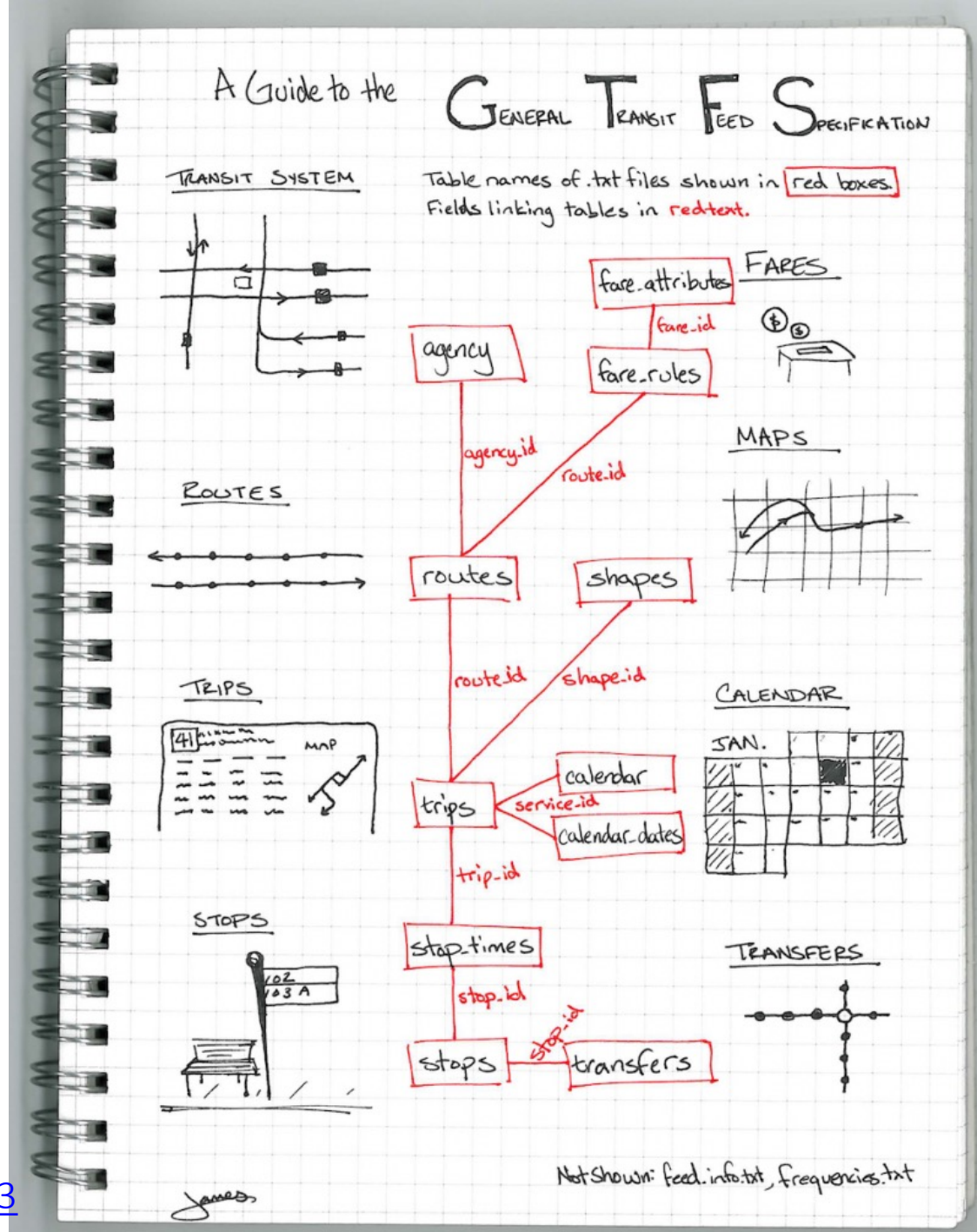
General Transit Feed Specification (GTFS)

Qualche esempio:

[Percorsi Amat Palermo]

https://transit.land/

http://tracker.geops.ch/



https://opendata.comune.palermo.it/opendata-dataset.php?dataset=1303

Statistical Data and Metadata eXchange (SDMX)

Lo standard SDMX (Statistical Data and Metadata eXchange) è un linguaggio XML per lo scambio di dati e metadati statistici. L'iniziativa è promossa e incoraggiata da istituzioni internazionali tra le quali la Bank for International Settlements, la Banca centrale europea, l'Ufficio statistico della comunità europea (Eurostat), il Fondo monetario internazionale, l'OCSE, le Nazioni Unite (Statistics Division) e la Banca Mondiale.

Lo standard include anche specifiche addizionali (ad es. sui web service). Lo standard SDMX è stato riconosciuto come uno standard ISO (ISO/Technical Specification 17369:2005).

http://ec.europa.eu/eurostat/web/sdmx-web-services/example-queries

Formati per dati meteorologici

Il Regolamento UE per i dati metereologici suggerisce l'utilizzo dei seguenti formati:

- **JSON** per dati orari;

- **BUFR** (Binary Universal Form for the Representation of meteorological data), formato di dati gestito dall'Organizzazione Meteorologica Mondiale (WMO – World Meteorological Organization);

- **NetCDF** (Network Common Data Form), insieme di librerie software e formati di dati indipendenti dalla macchina che supportano la creazione, l'accesso e la condivisione di dati scientifici array-oriented

Per i dati del modello **NWP** (Numerical weather prediction), oltre al JSON e a NetCDF, si può utilizzare il formato **GRIB** (General Representation of fields In Binary), rappresentazione binaria di dati risultanti da un'osservazione o da una simulazione su modello numerico di una proprietà osservabile in un dominio spaziale e temporale su un sistema di riferimento geospaziale o celeste.

Parquet

- formato open-source di file di dati a colonne

- pensato per lo stoccaggio e il recupero efficiente dei dati

- Elevata efficienza nella compressione e decompressione di dati.

- Esempi:

https://opencoesione.gov.it/it/opendata/#!progetti_section