FAIR O NON FAIR? QUESTO (NON È) IL PROBLEMA… DEI DATI

TESSA PIAZZINI
BIBLIOTECA BIOMEDICA
UNIVERSITÁ DEGLI STUDI DI FIRENZE

COMINCIAMO DA QUI… CORREVA L'ANNO 1996

- L'Human Genome Project e i Bermuda Principles:
  John Sulston e Bob Waterston chiedono di fissare delle regole di base per gli scienziati coinvolti nella mappatura del genoma umano

1. All parts of the human genome sequenced during the Human Genome Project should be distributed into the public domain.

2. DNA sequences had to be released rapidly into the public domain within 24 hours of completion.

3. Sequencing centres should inform the Human Genome Organisation (HUGO) of their intentions to sequence certain regions of the genome so that they could present the information online.

Fonte: Yourgenome.org, 2016, **«How did the Human Genome Project make science more accessible?"**

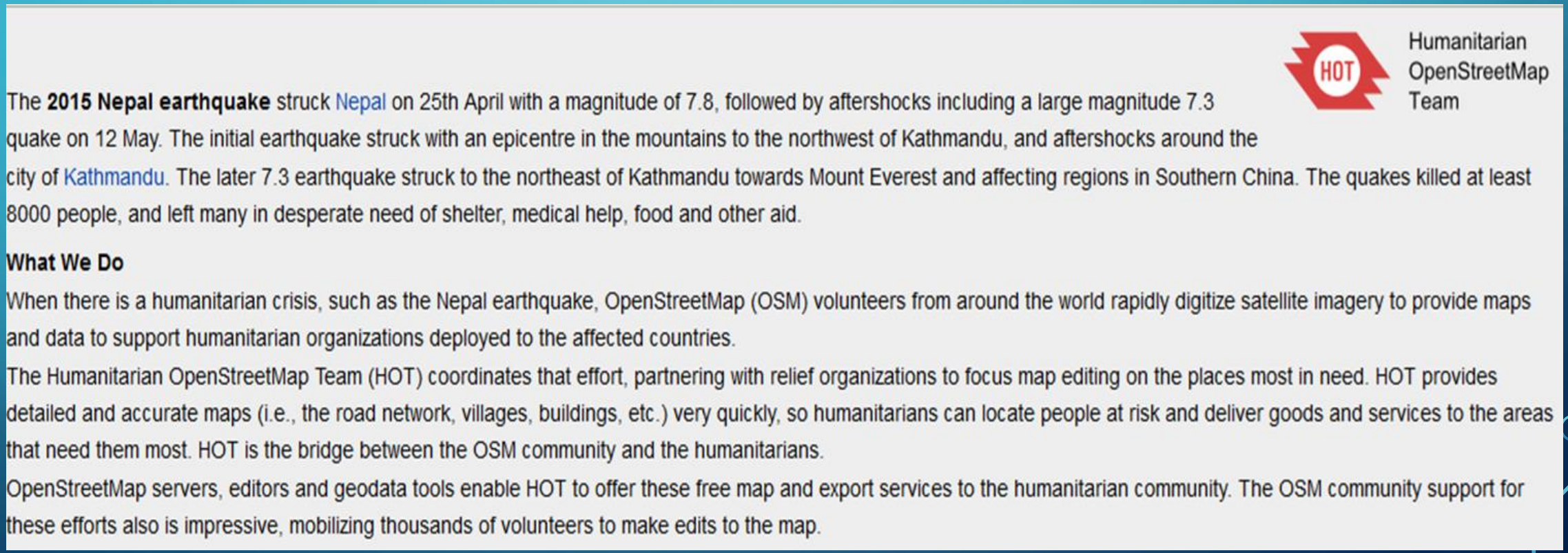
DIECI ANNI DOPO…

The global spread of the H5N1 avian influenza virus has already extensively damaged economies worldwide and food safety in developing countries. The spread of infection to new ecosystems results in adaptation of the virus to new hosts, including humans, which amplifies the potential for a flu pandemic. Because it is recognized that avian influenza viruses may be the progenitors of the next human pandemic virus, their genetic evolution should be tracked in detail and promptly investigated.

Several countries and international agencies have recently taken steps to improve sharing of influenza data[1,2,3,4], following the initiative of leading veterinary virologists in the field of avian influenza. The current level of collection and sharing of data is inadequate, however, given the magnitude of the threat. We propose to expand and complement existing efforts

Fonte: Peter Bogner, Ilaria Capua, David J. Lipman, Nancy J. Cox & others, «**A global initiative on sharing avian flu data**», *Nature* **vol. 442**, p. 981(2006)

DI DIECI ANNI IN DIECI ANNI (QUASI)



Humanitarian OpenStreetMap Team

The **2015 Nepal earthquake** struck Nepal on 25th April with a magnitude of 7.8, followed by aftershocks including a large magnitude 7.3 quake on 12 May. The initial earthquake struck with an epicentre in the mountains to the northwest of Kathmandu, and aftershocks around the city of Kathmandu. The later 7.3 earthquake struck to the northeast of Kathmandu towards Mount Everest and affecting regions in Southern China. The quakes killed at least 8000 people, and left many in desperate need of shelter, medical help, food and other aid.

What We Do

When there is a humanitarian crisis, such as the Nepal earthquake, OpenStreetMap (OSM) volunteers from around the world rapidly digitize satellite imagery to provide maps and data to support humanitarian organizations deployed to the affected countries.

The Humanitarian OpenStreetMap Team (HOT) coordinates that effort, partnering with relief organizations to focus map editing on the places most in need. HOT provides detailed and accurate maps (i.e., the road network, villages, buildings, etc.) very quickly, so humanitarians can locate people at risk and deliver goods and services to the areas that need them most. HOT is the bridge between the OSM community and the humanitarians.

OpenStreetMap servers, editors and geodata tools enable HOT to offer these free map and export services to the humanitarian community. The OSM community support for these efforts also is impressive, mobilizing thousands of volunteers to make edits to the map.

Fonte: OpenStreetMap Wiki

E ARRIVIAMO A OGGI…

- L'emergenza COVID ha mostrato, per l'ennesima volta, quanto sia necessario condividere i dati, eppure ancora…

Tuttavia, se da un lato la pandemia ha evidenziato ulteriormente l'importanza dei dati, dall'altro ci ha mostrato che l'UE non era pronta a sfruttare appieno il loro potenziale. La situazione di crisi ha amplificato problemi pregressi. Per esempio, la limitata interoperabilità dei dati sanitari all'interno e tra Stati Membri ha ostacolato la condivisione di tali dati cruciali.

Fonte: Maria Rosaria Coduti, «"I dati risorsa cruciale, la lezione del covid-19 per il futuro dell'Europa", su Agenda Digitale, 31 agosto 2020

Se ci fosse già stato uno spazio comune europeo di dati sanitari, le autorità sanitarie avrebbero potuto reagire prontamente ed avrebbero avuto gli strumenti per utilizzare i dati come una risorsa essenziale per fronteggiare il virus.

5 October

Nearly 16,000 coronavirus cases were missed in the UK due to a technical glitch

Almost 16,000 coronavirus cases in England were missed from official daily UK case figures between 25 September and 2 October due to a technical mistake, according to Public Health England. A total of 15,841 cases were left out of the daily UK figures over the eight-day period, or about 1980 missed cases per day. The missing cases were added over the weekend, artificially raising daily UK case numbers to 12,872 for Saturday and 22,961 for Sunday. Public Health England was reportedly using Microsoft Excel software as a makeshift database to record lab cases. The file reached the maximum number of columns, which cut off thousands of cases.

"Some of the data, it got truncated and it was lost," UK prime minister Boris Johnson told journalists

Fonte: https://www.newscientist.com/article/2237475-covid-19-news-one-in-170-people-in-england-have-

QUALCOSA ANCORA NON STA FUNZIONANDO

- L’impatto di una errata o carente o assente gestione dei dati ha enormi ripercussioni anche sulle buone pratiche di condivisione dei risultati della ricerca.
- Secondo uno studio recente l’emergenza Covid ha generato un’esplosione dei preprint o di pubblicazioni OA, ma spesso senza dataset allegati o metodologia utilizzata
- Risultato? Un alto numero di ritrattazioni e In fondo «*Una ricerca pubblicata senza dati è solo «la pubblicità della ricerca»» (*Barkhiet, Donoho 1995)
- confusione (allarmismo o facili speranze) nella collettività, perché presentati dalla stampa come risultati definitivi.

EPPUR (QUALCOSA) SI MUOVE…

- Le politiche europee stanno spingendo molto sull'importanza della corretta gestione dei dati, sia per i benefici economici (Mercato Unico Digitale) che per quelli sociali.
- Aprile 2018: Commissione Europea ha aggiornato le precedenti raccomandazioni (417/2012) sull'accesso all'informazione scientifica e sulla sua conservazione.
- Se il titolo delle raccomandazioni è rimasto inalterato una lettura comparativa delle due versioni mostra l'evidente cambio di passo e una maggiore consapevolezza della complessità delle modalità di produzione scientifica.

E INFATTI… TROVA LE DIFFERENZE

- «L’impegno per la progressiva introduzione dell’accesso aperto è di portata mondiale»
- «Accesso aperto ai dati di ricerca»
- «i dati di ricerca prodotti nell’ambito di attività di ricerca finanziate con fondi pubblici siano pubblicamente accessibili, utilizzabili e riutilizzabili per mezzo di infrastrutture elettroniche digitali»

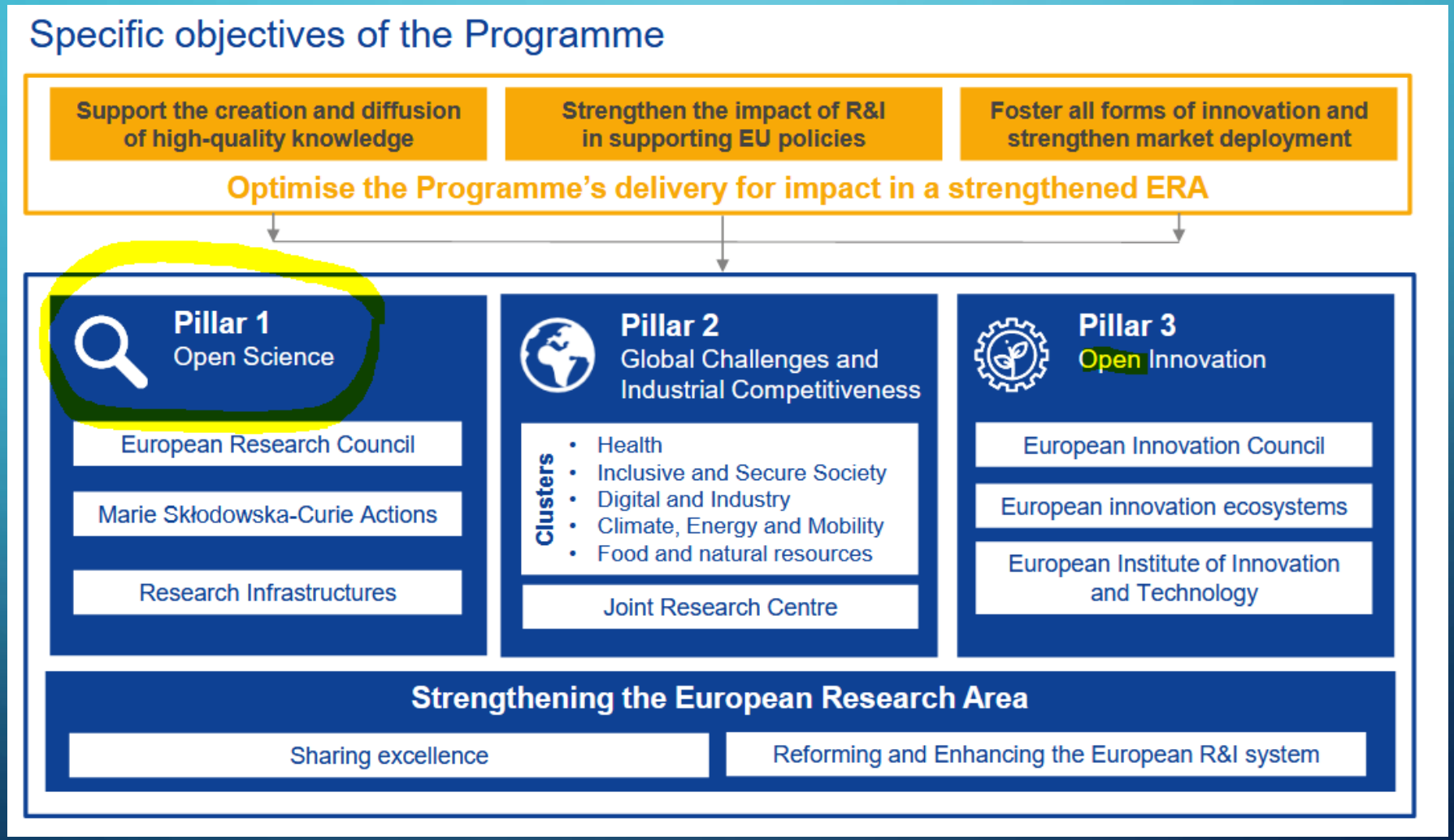
- Da «Infrastrutture elettroniche»

- «Il movimento verso l'accesso aperto è una tendenza mondiale»
- «Gestione dei dati di ricerca, compreso l'accesso aperto»
- «i dati di ricerca prodotti nell'ambito di attività di ricerca finanziate con fondi pubblici diventino e rimangano reperibili, accessibili, interoperabili e riutilizzabili («principi FAIR») in un ambiente sicuro e affidabile, per mezzo di infrastrutture digitali (comprese quelle aggregate nell'ambito del cloud europeo per la scienza aperta, se del caso), salvo che questo non sia possibile o sia incompatibile con l'ulteriore sfruttamento dei risultati delle attività di ricerca («il più aperto possibile e chiuso solo quanto necessario»)
- «Infrastrutture per la scienza aperta»

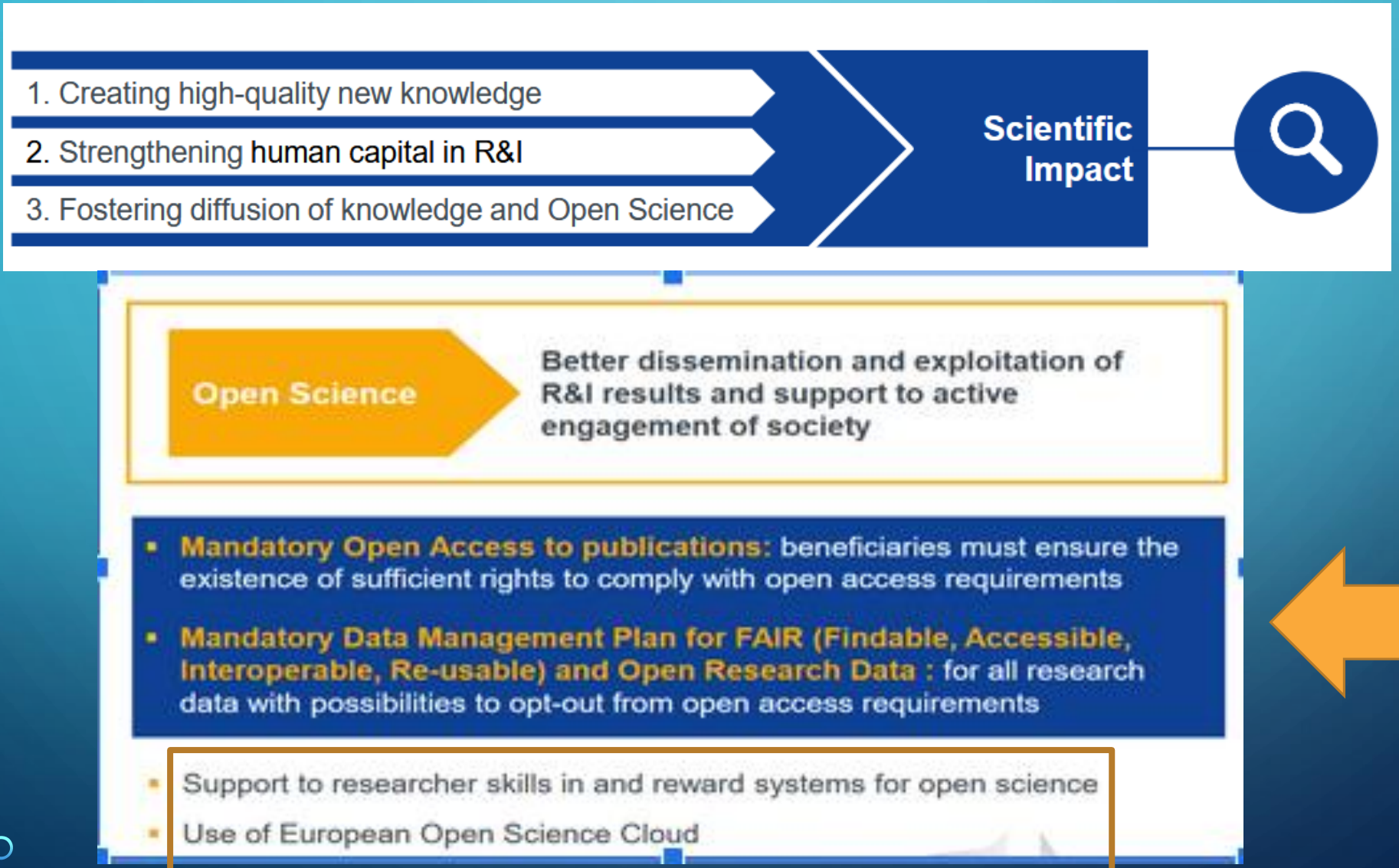
Due nuove sezioni dedicate a «Capacità e competenze» e «Incentivi e ricompense»

Lo spostamento di asse del «fare ricerca» non può passare solo dalla buona volontà dei singoli, dalla spinta etica e morale degli individui, né dall’obbligatorietà, ma anche da un adeguamento dei criteri valutativi, di promozione e di avanzamento di carriera.

OPEN SCIENCE NEL PROSSIMO HORIZON EUROPE



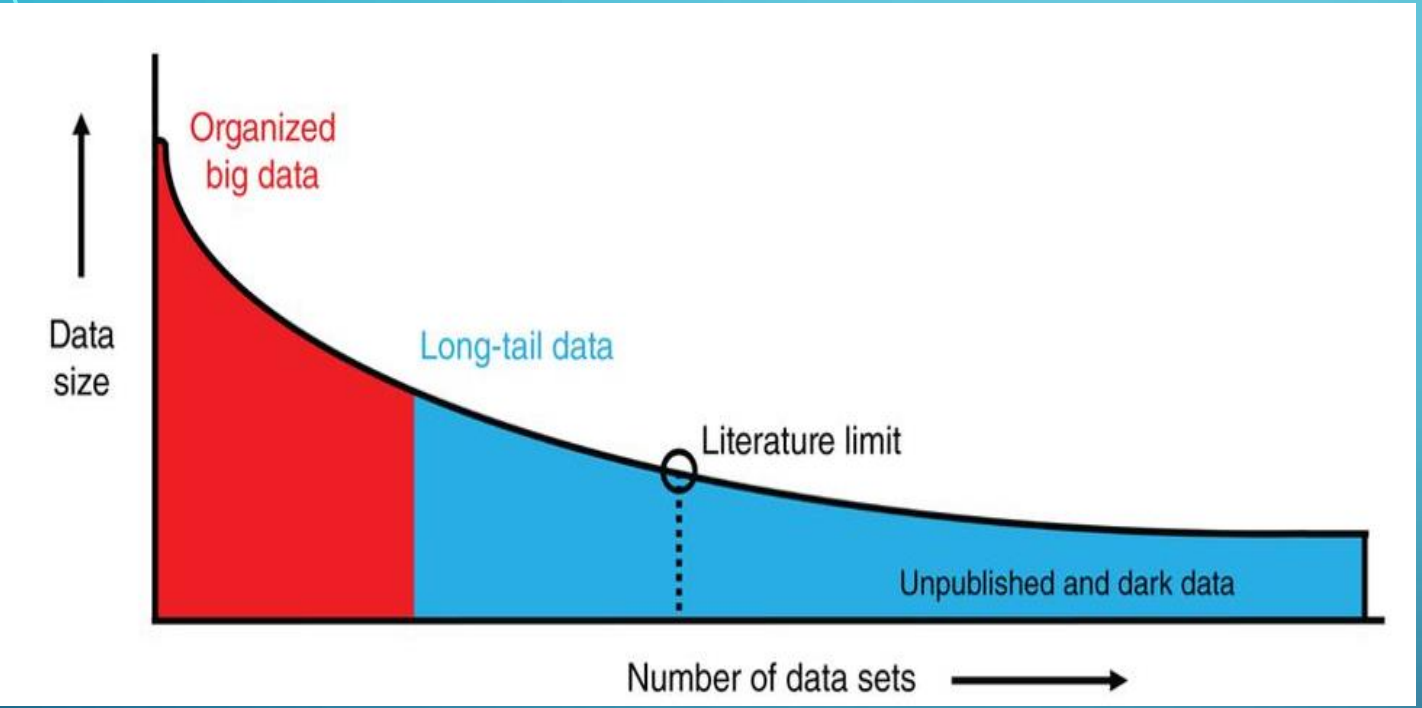
SUL PIANO CONCRETO: LE RICHIESTE



DALLA TEORIA ALLA PRATICA

- Maggiore sarà la corretta gestione dei dati di ricerca maggiore sarà il riverbero benefico a livello di comunità scientifica, collettività e competitività economica.
- Mentre la realizzazione dell'accesso aperto alle pubblicazioni non ha motivazioni endogene alle pubblicazioni che la possano frenare, nel caso dei dati di ricerca prima di parlare di apertura dobbiamo concentrarci su una gestione accurata, basata su due elementi sostanziali: la redazione di un Data Management Plan (DMP) e le attività necessarie per rendere i dati FAIR

IL PROBLEMA DELLA «LUNGA CODA»



Fonte: Big data from small data: data-sharing in the 'long tail' of neuroscience, *Nature Neuroscience* **volume 17**, pages 1442–1447 (2014)

A ogni dato il suo problema: per i Big data fondamentalmente la conservazione su lungo periodo, ma sono spesso dati omogenei, standardizzati e regolati.
Per la «lunga coda» il problema principale è il riuscire a trovare o a utilizzare i dataset, spesso relativamente piccoli, eterogenei e nascosti nei computer dei singoli ricercatori; oppure pubblicati in formati che non permettono il text e data mining.

La regola aurea è «aperti il più possibile, chiusi quando necessario», tenendo quindi conto di restrizioni di carattere commerciale, etico, e di privacy per i dati personali e sensibili, ma garantendo che siano accessibili, ricercabili, utilizzabili, valutabili e comprensibili: in una parola … **FAIR**

DMP: NON UN DOCUMENTO, MA UNO STRUMENTO DI LAVORO

- Il DMP è uno strumento in evoluzione che segue tutto il progetto di ricerca e lo accompagna in ogni sua fase, e che rende conto di **tutto** il ciclo di vita dei dati
  (data collection, data documentation, data storage & back up, data access & security, data preservation & reuse)
- Ci sono ormai numerosi strumenti e template già predisposti per la redazione di un DMP
  - ARGOS
  - DCC DMPonline Tool
  - Griglia IOSSG (Italian Open Science Support Group)

IL CUORE DI TUTTO: RENDERE I DATI FAIR

"The word play with '**fairness**', in the sense of **equity** and **justice**, has also been eloquent in communicating the idea that FAIR data serves the best interests of the research community and the advancement of science as a public enterprise that benefits society"

European Commission, Turning FAIR into reality: final report, 2018

- **Findable:** reperibili, facili da trovare per persone e macchine (PIDs e metadati)
- **Accessible:** attraverso un protocollo di comunicazione standard e open, possibilmente che includa autenticazione o autorizzazione. I metadati dovrebbero rimanere sempre accessibili. **Accessibili non vuole dire necessariamente «aperti»**
- **Interoperable:** combinabili con altri dati o strumenti, quindi il **formato** dovrebbe essere open
- **Reusable:** ben descritti per permetterne il riutilizzo e forniti di una licenza chiara e accessibile.

MA COME SI FA?

The FAIR guiding principles: https://doi.org/10.1038/sdata.2016.18

To be Findable:

F1. (meta)data are assigned a globally unique and persistent identifier

F2. data are described with rich metadata (defined by R1 below)

F3. metadata clearly and explicitly include the identifier of the data it describes

F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1. the protocol is free, open and universally implementable

A1.2. the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

I2. (meta)data uses vocabularies that follow FAIR principles

I3. (meta)data include qualified references to other (meta)data

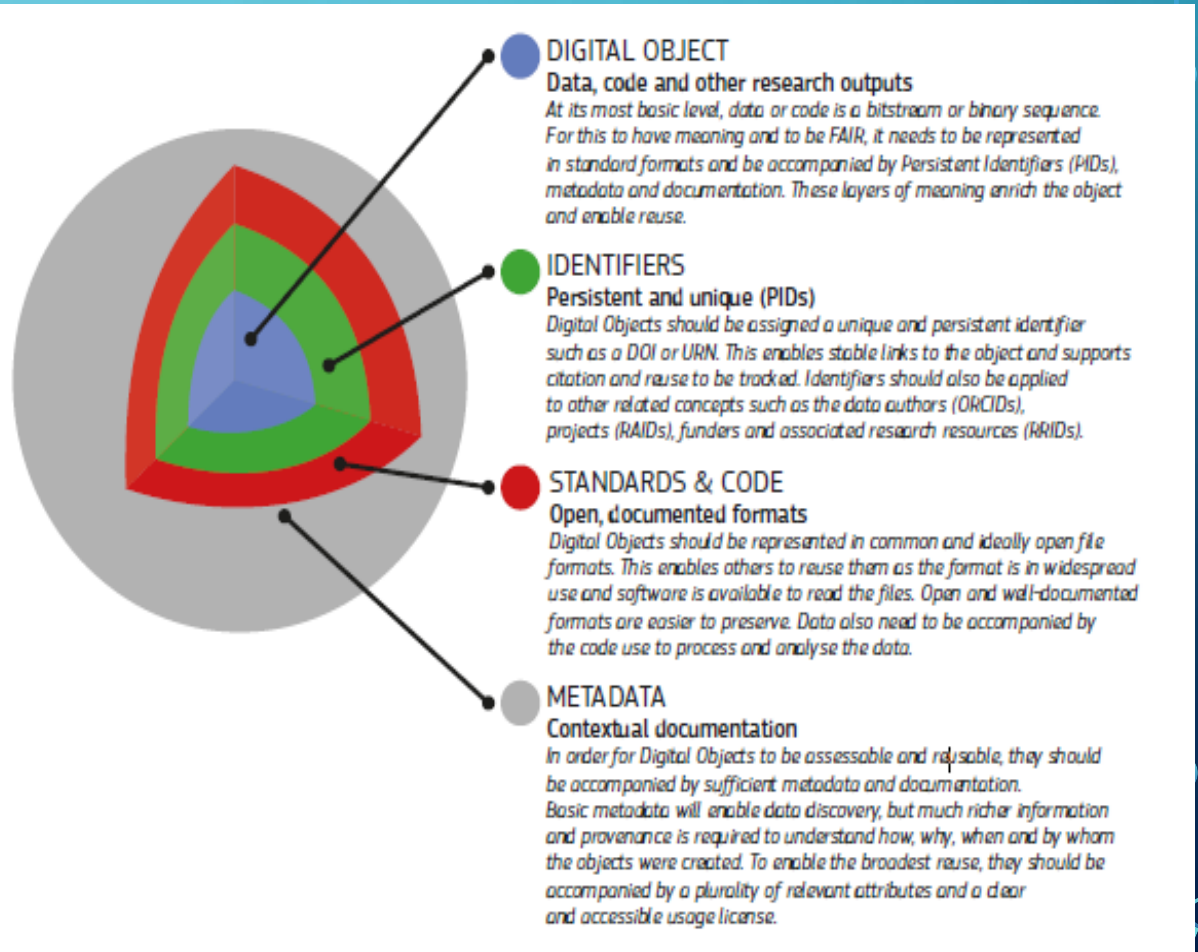
To be reusable:

R1. (meta)data are richly described with a plurality of accurate and relevant attributes

R1.1. (meta)data are released with a clear and accessible data usage license

R1.2. (meta)data are associated with data provenance

R1.3. (meta)data meet domain relevant community standards



European Commission, Turning FAIR into reality: final report, 2018

- Un’analisi dettagliata è offerta dalla GoFAIR Initiative o da OpenAIRE, ma sono numerose le realtà che offrono strumenti per guidare i ricercatori nella costruzione di dati FAIR:
- Checklist CESSDA
- Checklist EUDAT
  ARDC Self Assessment Tool
- FAIRDOM Checklist

“Data-related aspects need to be taken into account from the earliest project stage and fully incorporated in project plans, funding requests and reporting. Taking the research workflow as a sequence of stages, **FAIRness needs to be considered throughout the research process in the form of a set of questions relating to data**, which should be a basis to define the data management plan”

Fonte: *European Commission, Turning FAIR into reality: final report, 2018*

COME SI MANGIA UN ELEFANTE?