opencare

Data strategy

Alberto Cottica – Edgeryders

This version: June 2016

This document is published under a Creative Commons Attribution 4.0 International license

SUMMARY

- Vision
- Generating high quality data
- Storage, sharing, metadata, preservation, license.
- Other data

Vision

photo: Andreas Overland

What are data for?

- “Learn-by-doing how to deploy collective intelligence to design care services” (Objective 1)
- “Assemble a software stack to monitor and assist collective intelligence social dynamics in online communities” (Objective 3)
- Data analysis in opencare is meant to **observe and understand collective intelligence in action**.

Semantic Social Network Analysis

- Open online conversations encode collective intelligence
- We represent them as networks of social interactions that carry meaning (semantics)

CAPS vs. SSNA

Collective → Social

Collective intelligence is **interactional**, not additive.

Awareness → Semantic

Interactions encode meaning. Collective intelligence is used for **sensemaking and scenario exploration**, not computation.

Platforms → Network

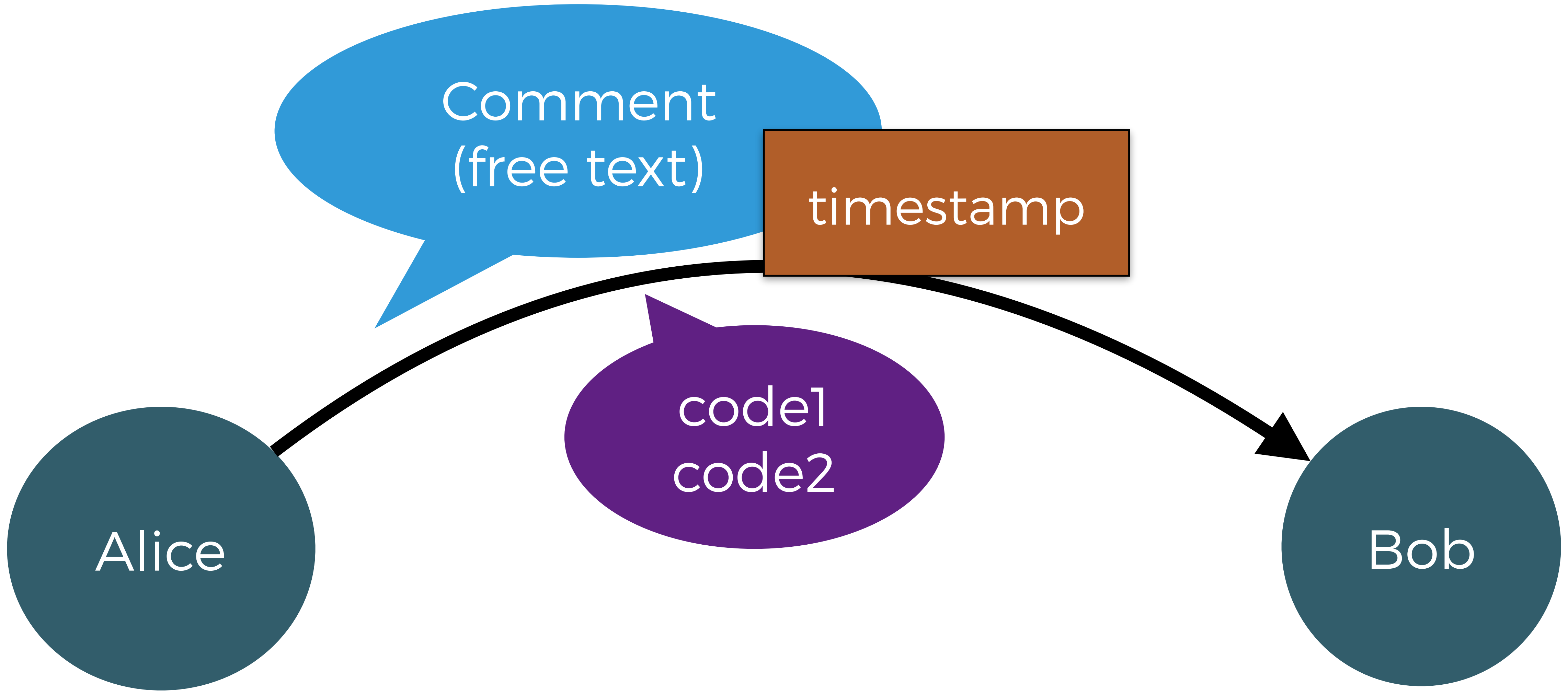
Interaction happens on digital platforms and is represented as a multilayer graph.

A DATA MODEL FOR SSNA

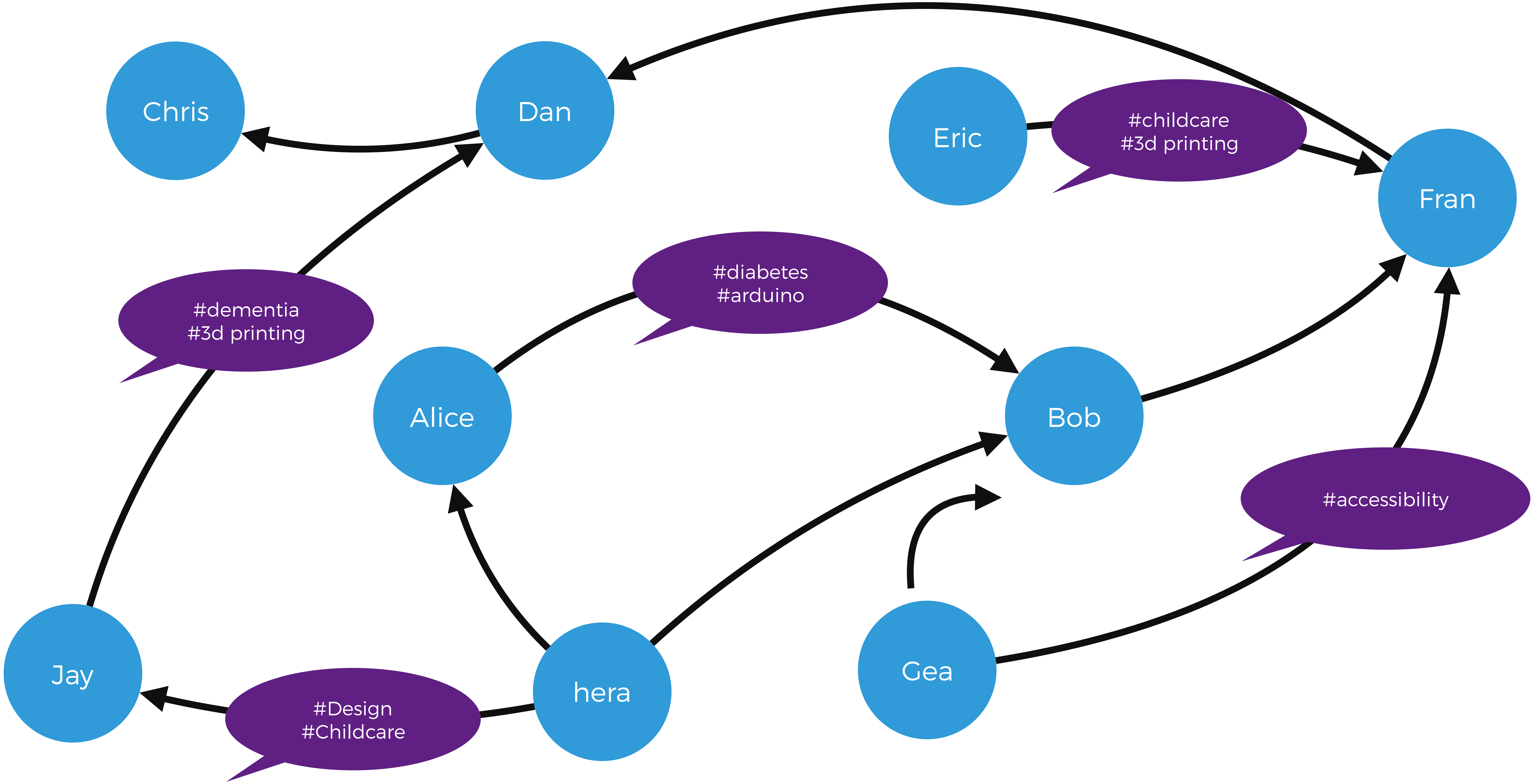
- We call **contribution** the atomic component of a Semantic Social Network. It can be standalone (a post/ thread opener) or a response (a comment/thread reply).

- Contributions have: **source** (author) and **target** (addressee) (inducing the social network), **timestamp** (dynamics), and **text**. Standalone contributions are intended to address the whole community, and as such have no target.

- Contributions are primary data. Text is later enriched by ethnographic **codes**, opencare's secondary data, and acquires semantics.



posts/comments induce network interaction



CONTRIBUTIONS COMBINE INTO A SEMANTIC NETWORK

WE CAN THEN STUDY IT AS A MATHEMATICAL OBJECT.

photo: Edgeryders

Generating high quality data

THE "PERFECT" PRIMARY DATA

- Ideal scenario: Anna writes a contribution directly in the Edgeryders platform, in English.
- Low collection cost: the contribution is immediately available, human- and machine-readable.
- No empty fields and missing values: platform preserves source, target, timestamp.
- No intermediary between contributor and text: the contribution is equivalent to the transcription of a field interview in (offline) ethnography.
- Immediately available for future interaction.

Primary data Quality issues and mitigation

Issue	Mitigation
Alice makes a contribution in the context of an offline event.	Ask Alice to write a challenge response on the platform summarising her thoughts.
Bob is not comfortable communicating in writing.	Interview Bob, Then create an account to his name (or a pseudonym) with his email and upload his contribution onto the platform. Try to use Bob's own words rather than your own, Future comments will automatically be notified to him, encouraging interaction.
Chris wants to make a contribution, but does not want to give her name.	Explain Chris that there is no need to give her real name. Account creation only needs an email, even a one-time address will do. Email addresses are stored under strong encryption, and login is also encrypted (https).
Dan feels his English is not good enough.	It's perfectly OK for Dan to write in his own native language. Ethnographic coding will be intermediated by translation, by humans if possible, by machines otherwise.

Secondary data: ethnographic coding

- A professional ethnographer reads the material and assigns keywords, known as “codes”, to snippets of text.

- Codes are normally arranged in a hierarchy: for example “Germany” might be a code belonging to the “Places” code class.

- Codes are generated and managed through an application called Open Ethnographer (https://edgeryders.eu/en/open-ethnographer). The edgeryders.eu platform stores codes into the same database as the primary data.

Tertiary data: semantic social network

- The primary and secondary data are combined to generate a semantic social network.
- The full network is stored separately from the primary and secondary data.

STORAGE, SHARING, METADATA PRESERVATION, LICENSE

photo: James West

TYPES OF DATA

- **Primary data**. Contributions as formed on the edgeryders.eu platform (users, posts, comments)

- **Secondary data**. Ethnographic codes on the free text fields of the primary data.

- **Tertiary data**. Primary and secondary data rearranged as a semantic social network.

Primary and secondary data

- Storage: edgeryders.eu platform, secured by https
- Sharing (2016-2017): through APIs (Views JSON Drupal module) that return a JSON file containing fresh data. Users are identified by numeric IDs, not by name.
- Metadata: via API documentation on www.mashape.com (large, active community, developer-friendly documentation). Redundancy on edgeryers.eu as a wiki. We use W3C's Guidelines for best practice: https://www.w3.org/TR/dwbp/#accessAPIs
- Sharing and conservation (from late 2017): at the end of the project, the latest version of the JSON files are dumped onto zenodo.org. The metadata of the data dump will be documented using the Data Package standard (http://specs.frictionlessdata.io/data-packages/)

Tertiary data

- The complete graph is maintained in a Neo4j graph database on the Edgeryders server.

- Periodic calls to the edgeryders.eu APIs add new posts, comments and codes to the graph database.

- Tertiary data will be shared in the late stages of the project, in the form of a Neo4j data dump hosted on zenodo.org. Metadata will be made available in the Data Package standard (http://specs.frictionlessdata.io/data-packages/), and more specifically its extension to graph data developed in the course of an FP7 project called CATALYST (https://github.com/Wikitalia/edgesense/tree/master/documentation/Example%20Datapackage). This solution also takes care of long term preservation.

License

- Primary, secondary and tertiary data as defined in this document are published under a **Creative Commons Attribution 4.0 International** license: https://creativecommons.org/licenses/by/4.0/

How our data strategy ensures:

- **Discoverability**. Mashape and Zenodo both assign unique identifiers to, respectively, API documentation and datasets. Both have large community of reusers.
- **Accessibility**. Good documentation and open license make the data simple to reuse.
- **Intelligibility**. Good documentation is provided.
- **Usefulness for purposes different than the one for which the data were collected**. We already foresee possible reuses in areas like NLP and machine learning.
- **Interoperability to standards**. Wherever possible (data dumps), we use Open Knowledge standards. Where this is not possible (API documentation), we fall back on the wisdom of the Mashape community.

Other data

- opencare is expected to generate data in forms that do not lend themselves to interpretation as a semantic social network.
- Examples: vectorial files generated in the lab under WP3; field notes from Task 1.4; survey data under Task 4.2.
- As those activities progress, our data management plan will be expanded to cover those types of data too.