Data Observatories

Data observatory is a navigation apparatus which can be used to orient oneself in the vast landscape of data on radio transmissions based on computable similarity. The development of visual instruments, *data observatories*, is part of the method and ambition to organize data on radio signals according to properties of the dataset.

The dataset used by the *data observatory* comes from the Signal Identification Guide (SIGID) wiki: an organized collection of information about radio signals, held among a community of radio amateurs and enthusiasts. As of March 2022, there were 444 known or identified and 328 non-identified signal pages on the website. Known signals are divided into categories based on different listener community interests, such as the military, amateur radio, trunked signals or satellite reception. Each signal is characterized by its transmission properties, a short description and a recording sample with the spectrogram. Any radio signal that can be received and recorded can be included in the database.

The data itself mean very little in terms of human communication: it consists of demodulated recordings of radio transmissions pertaining to document the protocols rather than the content of transmitted messages. What was said is often a simple 'I am there', sent out by radio beacons, industrial and monitoring equipment. Partial access to this data means extracting specific features from the data, which gives access to a comparison across the dataset.

To render them comparable, or commensurable, a self-organizing map (SOM) machine learning algorithm was used to encode signals in terms of specific properties such as probability of silence, the level of noise in the audio sample, or an audio identification technique called

fingerprinting.[1] Networks of machine learning algorithm train on each of these property sets and produce an organized space – a grid of 'codebook vectors'[2] – that can be navigated and explored in three dimensions: according to proximity of codebook cells (horizontally and vertically) as well as according to the content of one cell (depth). By doing this, the database gets an informational face, a different one for each of the properties. Signals are grouped with other signals in one cell of the grid (see Fig 01) when they exhibit strong similarity according to the organizing property. Signals that occupy opposing corners of the grid should in principle pertain to very different types, while types themselves are fluid across the grid – there is no hard cut to determine the area in which signals belong to any particular type.

The development of *data observatories* contributed to exploring ways to reorganize the digital archive on radio signals collected by radio enthusiasts and available online as Signal Identification Guide (SIGID) wiki.[3] This database documents listening practices of the community of radio amateurs and enthusiasts. In addition to radio amateurs, recordings of radio signals 'in the wild' can capture the interest of a telecommunications engineer and a media archaeologist, of a data scientist, and an ecologist, or a historian of science and technology.

The neural networks of the self-organizing map (SOM) lay down the ground, whose topology is rendered commensurable with an outside, intuitive domain: Descriptions and Projections.

---

1 Fingerprints are a sort of a condensed digital summary of an audio signal, based on peak points in the spectrogram which represent higher energy content. The technique is known for its use in Shazam music identification application. The related paper by Wang is available under https://zenodo.org/records/1416340, within a Shazam github repository: https://github.com/bmoquist/Shazam, and also within Columbia University repository of Dan Ellis, https://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf

2 A codebook vector is a list of numbers that have the same input and output attributes as the training data.

3 Signal Identification Guide wiki is available at http://sigidwiki.com (accessed 17.02.2022).