written by Fred van den Driessche on 8 January 2014 in Atlas, Atlas A-Z

# e is for equivalence

Continuing the Atlas A-Z series, this article is brought to you by the letter E. E is for Equivalence.

## what is equivalence?

In Atlas, data is ingested from a number of sources such as broadcasters, tv platforms and listings providers. Some sources might describe when a programme was broadcast and on what channel, some may list on-demand locations where you can catch-up with a show, while others might simply provide the topics covered by a particular episode. Atlas will often have data for a given programme from a number of different sources.

Equivalence is the sub-system in Atlas which determines that two pieces of data from different sources represent the same content, i.e. are in some sense equivalent, and links them together. For example, if you just missed a broadcast of hit motion picture The Bodyguard on Channel 4, Atlas could tell you it's available to watch on-demand on LoveFilm.

At the top level there are two major parts to Equivalence in Atlas. The first part is the generation of links between data points. The second is the resolution and merging of equivalence sets formed by those links.

## generating equivalences

The generation of equivalence links is currently a 5 step process for a piece of content (known as the subject):

1. **Candidate Search**: the first step is to find other data points that *might* be equivalent to the subject. This might be through broadcast co-occurrence (content broadcast simultaneously on the same channel is likely to be the same) or a simple title search.

2. **Candidate Scoring**: the candidates found are scored on a number of heuristics, for instance, title similarity, number of co-occurring broadcasts. A candidate can also be given a *null* score if a particular heuristic has no opinion on the similarity of the subject and candidate.

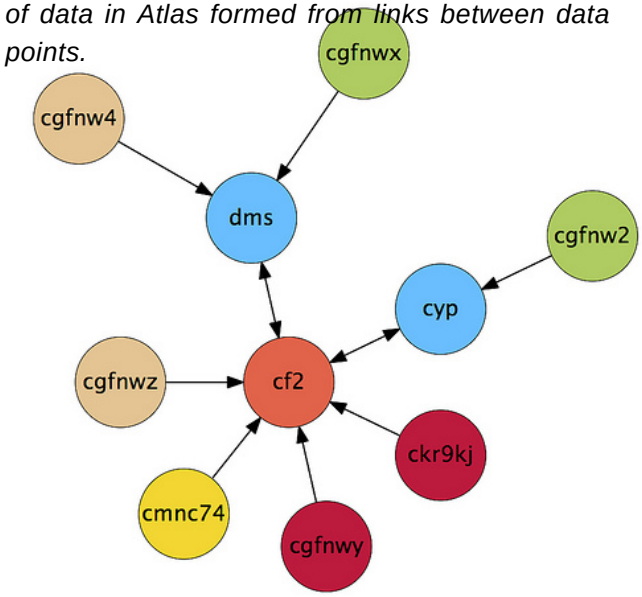
3. **Score Combining**: the scores from all the heuristics are combined for each candidate, currently this is a simple mean which ignores null scores.

4.

**Candidate Filtering**: a final backstop to ensure that no links are formed between data points which are known never to be equivalent.

5. **Equivalent Extraction**: the subject can only have one direct equivalent from each source so this step narrows down the field by determining the single strongest candidate per source.

**Figure 1**: *Graph representing an equivalence set of data in Atlas formed from links between data points.*




Equivalence in Atlas is an equivalence relation hence if a link is created marking A equivalent to B then B is
e equivalent to B, then it's

bine data points into sets, as
powerful as it means simple
t can't necessarily be linked

ked a set of data together and
nes through. What happens?
ey presented. There are two pertinent parts of the configuration:

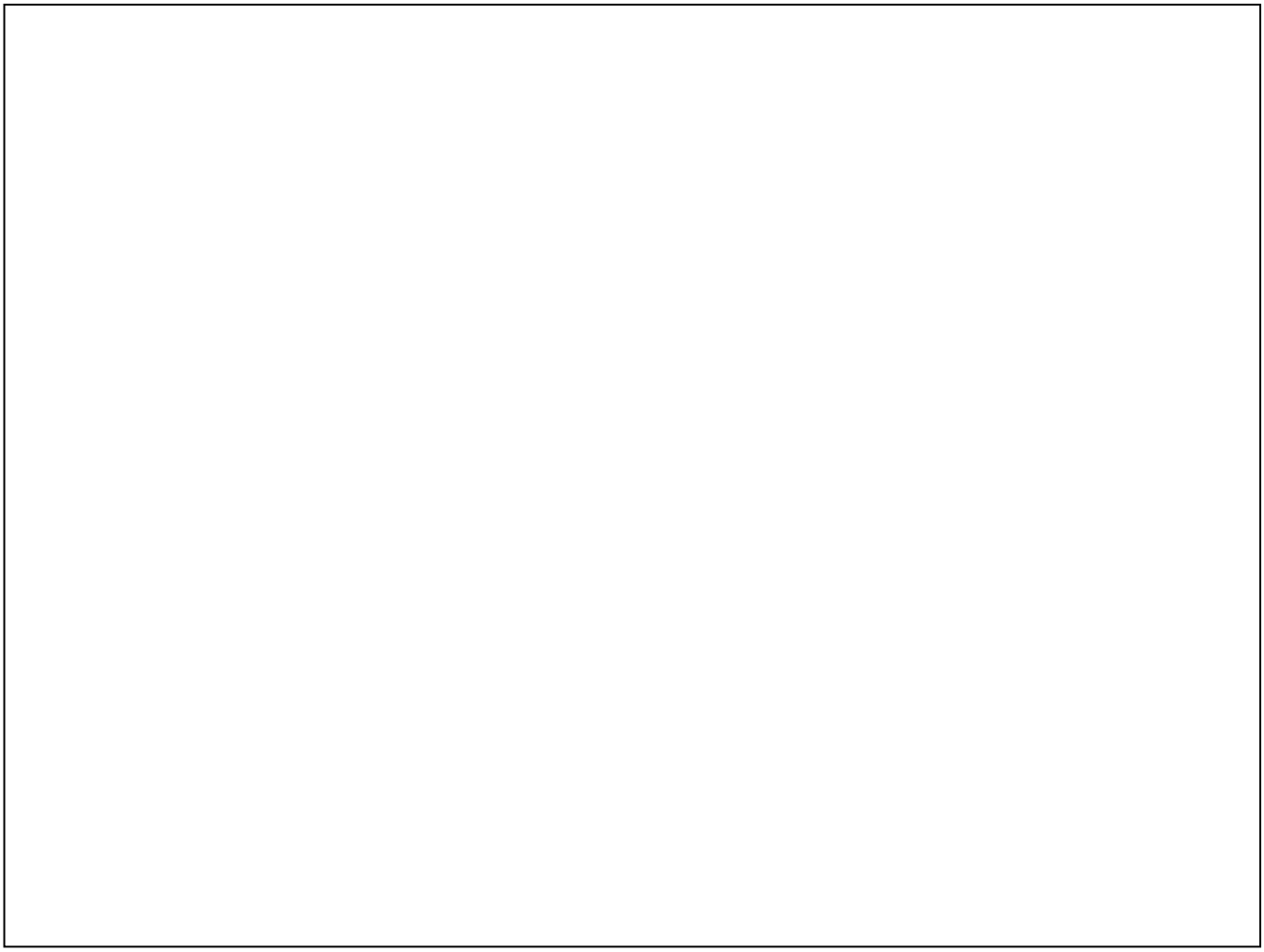
- **enabled sources**: only data points from enabled sources will be used to create the result.
- **precedence merging**: if enabled then the available data points will be merged into a single representation; otherwise the data points will be presented separately.

Put simply, the merging process will, for a field, work its way through the list of available sources and take the first non-null value available for that field. If the field is a collection then it will merge the values across all enabled sources and attempt to remove duplicates.

You can read more about this process in atlas: changes to id resolution and output.

## the bigger picture

The diagram below shows how the source data sets in Atlas are linked together by the Equivalence system (hover over each source to see its links to others.) Needless to say, these are very high level numbers, and mask all kinds of complexity about exactly what data is ingested from each source, and where it is and isn't linked. Many sources produce numerous feeds that are combined, and there are other sources where only a subset of data is currently ingested.



## more to be done

We're still learning and tweaking how equivalence works, it's an ongoing task since there are frequently new sources added with data characteristics which need to be catered for. We're also working on improving the internal modelling of equivalence graphs to improve resolution and merging. If you have any comments or questions please get in touch either below or via the Atlas discussion group.