

Concept Mapping for CSER's "TERRA" Project

Final Report

Version 1.0
haselwimmer@gmail.com

3rd April 2020

Contents

1. Overview.....	3
2. Description of concept map interface	4
3. Initial concept map – Raw Scopus data.....	7
Concept Map A1	7
Creating topic markers.....	9
Concept Map A2.....	12
4. Mapping manually-tagged XRisk terms	19
Concept Map B1	19
Concept Map B2	20
5. Comparing tagged and untagged articles.....	22
Concept Map A,B1	22
Concept Map A+,B1	23
Concept Map A+,B2.....	26
6. Excluding problem clusters	29
Concept Map R1	35
Conclusions.....	36
Appendix 1 - Description of concept mapping process.....	37
1. Text tokenization	37
2. Apply dimensionality reduction	37
3. Apply clustering to colour-code similar clusters	37

1. Overview

The **Existential Risk Research Assessment** (TERRA) project (terra.cser.ac.uk) is designed to create an existential risk bibliography using a combination of crowdsourcing and machine learning. **Concept mapping**, by contrast, provides a visual overview of text-based data in a way that provides insights into the structure of this data.

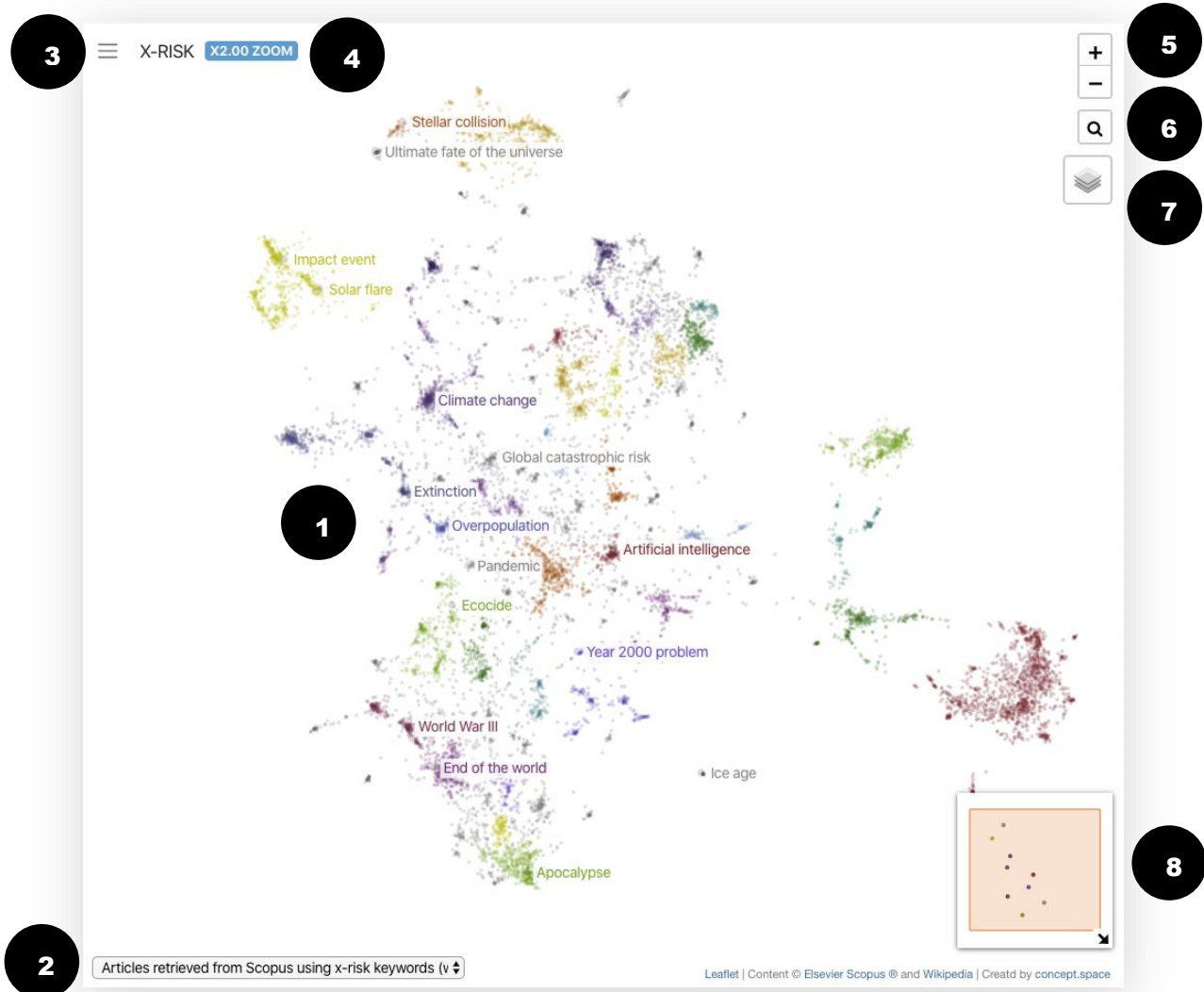
The purpose of this project is to use concept mapping to explore the text-based data of the TERRA existential risk bibliography. It is hoped concept mapping might provide crucial insights into the structure of TERRA's bibliography and might suggest ways to develop and refine the process of creating TERRA's existential risk bibliography.

The concept mapping implementation used in this project was developed by Stefan Haselwimmer (haselwimmer@gmail.com). An overview of how concept mapping works is outlined in the section "Appendix 1 - Description of concept mapping process". The text content used in the concept mapping was obtained from Elsevier's "Scopus" text library, as used in the TERRA system.

The interactive concept maps described in this document can be accessed at:

<https://terra.cser.ac.uk/map>

2. Description of concept map interface



1: Main area

The main area of the map shows the distribution of concepts and articles. At a distant view, many items will appear as coloured circles without titles - this is designed to avoid visual clutter, in a similar way to the way a country map will only show large cities. By zooming in closer and closer, a coloured circle becomes an active item with a visible title, and you can float your mouse over the item or click on the item to view more information.

2: View selector

A number of views have been pre-programmed into the interface and can be selected via a popup menu on the bottom-left of the screen. These views reference the map reference codes, eg. “**A1, B2**”, contained in this document.

3: Dynamic search

By clicking on the three bars on the top-left of the screen, you can enter a dynamic free text search of all articles on the map. Clicking on one of the articles will refocus the interface on that article. As you move about the map, the list of articles on the left-hand side of the screen will reload with articles “near” to the centre of the map.

4: Zoom scale indicator

The zoom scale indicator indicates the level of magnification of the current view.

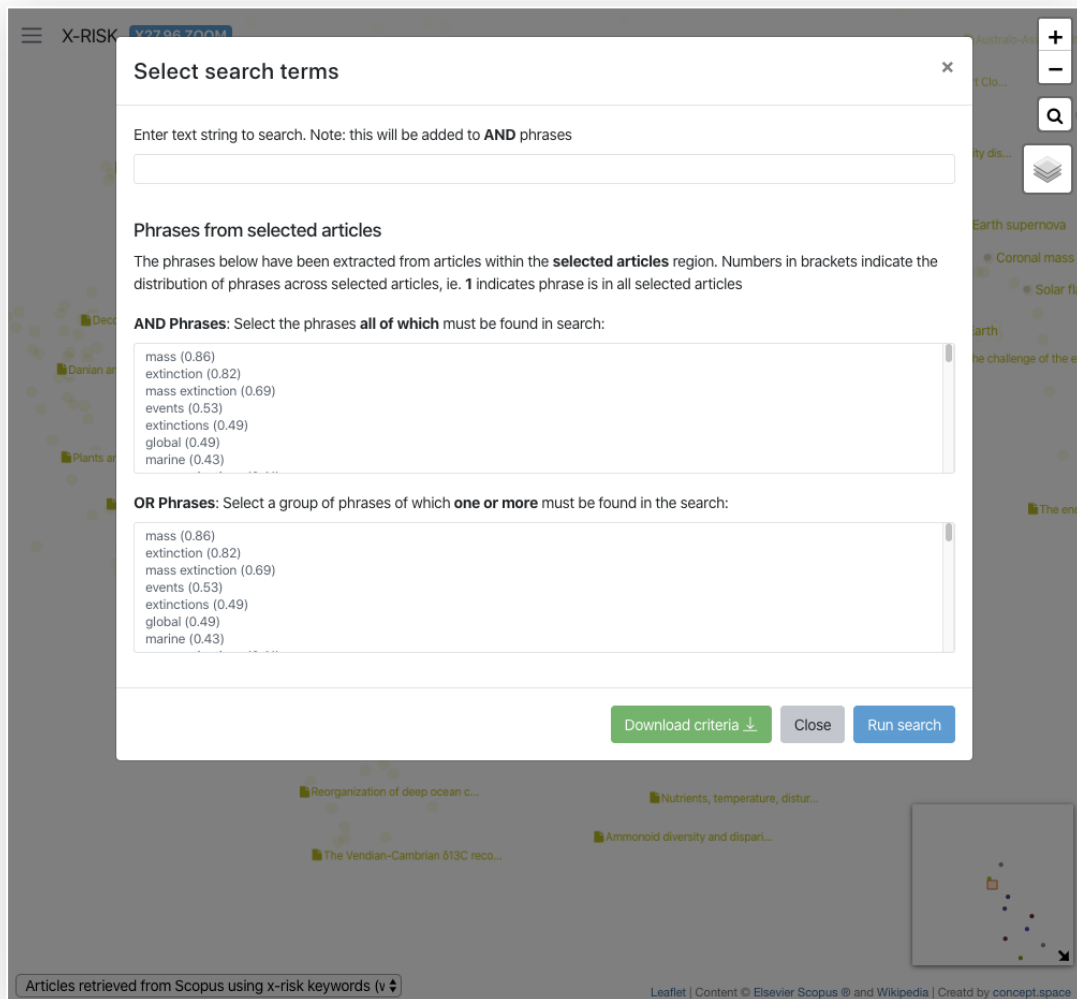
5: Zoom controls

Click on “+” to zoom in and “-“ to zoom out. You can also press the “+” and “-“ keys on your keyboard to zoom in/out.

6: Search and highlight

In addition to the free text search on the top-left of the screen, you can run a text search query across all articles (excluding concepts) and have the search results highlighted on the map. This is useful if you need to see the number of articles that have been retrieved from Scopus as a result of searching on a particular keyword or phrase.

The “Search and highlight” button can also be used to run queries on a combination of “AND” and “OR” keyphrases. These keyphrases are generated by shift-clicking and dragging a rectangle over a particular selection of articles – the system will analyse the selected articles within the rectangle and retrieve the most commonly-used keyphrases across the articles. For example:



7: Layer selector

The layer selector allows you to hide or show each of the three main layers:

- **Background:** The visual map of coloured circles, one for every article/concept.
- **Searched articles:** The articles retrieved from the “Search and highlight” function (see above).
- **Details:** The titles of items, visible when the user is sufficiently close. Hiding the “Details” layer may be useful if you need to see a broad overview of how items are distributed on the “Background” layer.

8: Overview

The “Overview” window gives you an overview of where you are on the overall map. An orange rectangle highlights the area you are viewing in the main window, and as you click-drag to move around the map, the orange rectangle will reposition itself accordingly. You can hide the overview window by clicking on the diagonal arrow, bottom-right.

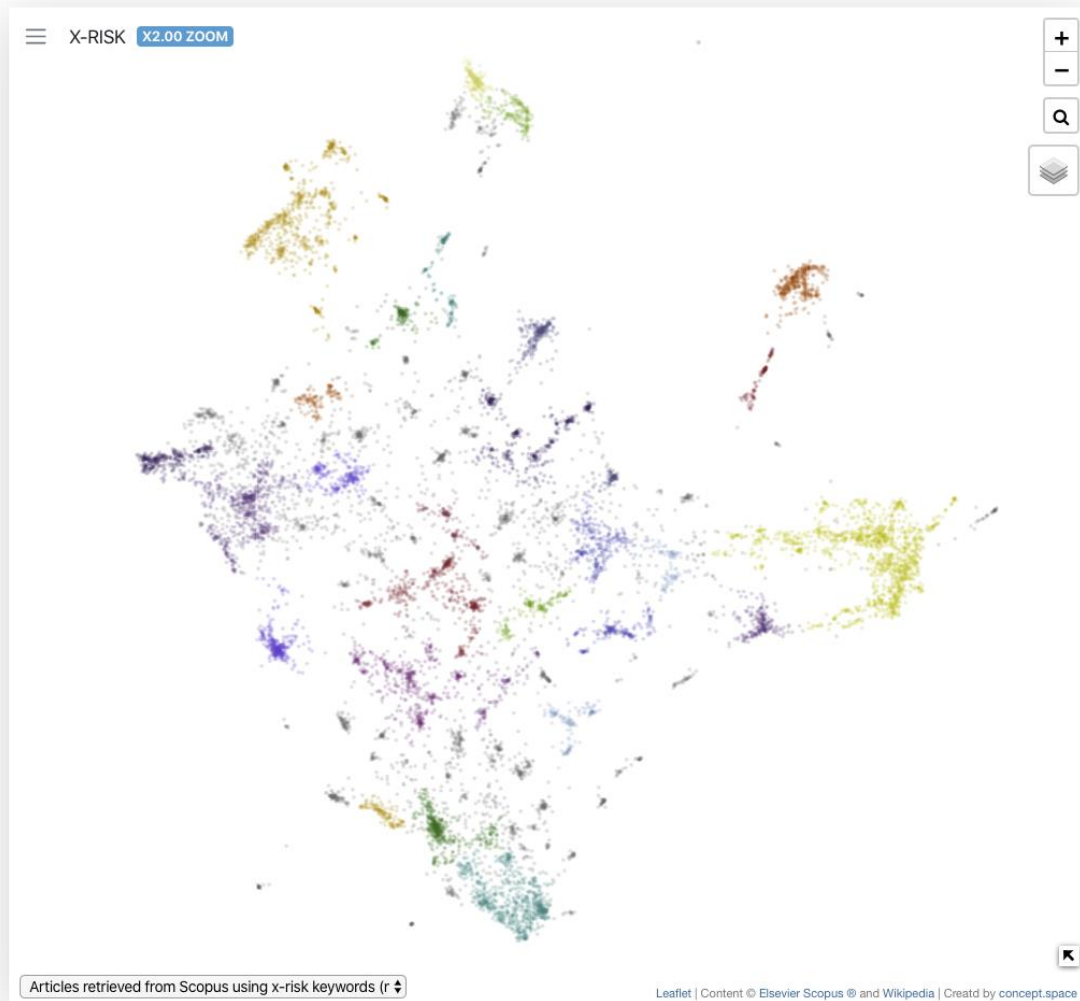
3. Initial concept map – Raw Scopus data

We took the database of raw search results from Elsevier’s Scopus data archive, as retrieved by the TERRA system, whose specification is set out here:

<https://www.sciencedirect.com/science/article/pii/S0016328719303702>

We then analysed these abstracts using concept mapping to create an initial concept map **A1**:

Concept Map A1



This concept map seems to reveal clear clusters. However, one should be wary of drawing hasty conclusions as some dimensionality reduction algorithms (used to create concept maps) artificially push data points together to improve the visibility of results.

Each coloured circle represents an article whose title will reveal itself once you zoom in sufficiently close. To zoom in or out, use the + or – buttons on screen or the scroll wheel on your mouse.

Zooming into some of these clusters show they are often focused on specific topic areas. If we zoom into the blue patch on the centre left for example, we get:



Creating topic markers

Zooming out, it will be difficult to see which topic a specific cluster relates to. It is therefore helpful to create “topic markers” that are visible when the view is zoomed out, eg. a visible marker for “climate change” in the middle of the climate change cluster that is visible at a low level of magnification.

To do this requires a canonical or typical-usage definition of each topic using the topic’s most commonly related terms. If the topic is “climate change”, for example, a canonical definition will refer to terms like “carbon dioxide”, “ice caps”, and “sea level rises”. When the concept mapping process is run on this canonical definition, it will be treated like a normal text item and – in theory at least – will be located close to other articles on climate change.

One way of providing a canonical definition is to use dictionary definitions of topics. Given the author’s previous experience with Wikipedia entries, it was decided to use Wikipedia to form the canonical definitions. One advantage of using Wikipedia is that entries are often very long and comprehensive, as compared to a typical dictionary definition.

The next question is how to form the list of topic markers for the existential risk concept map. We started with the list of topics that are used to query Scopus (Table 1: Scopus Search Terms). These are:

Table 1: Scopus Search Terms

Catastrophic risk, Existential risk, Existential catastrophe, Global catastrophe, Human extinction, Infinite risk, xrisk, x-risk, Apocalypse, Doomsday, Doom, Extinction of human, Extinction of the human, End of the world, World's end, World ending, End of civilization, Collapse of civilization, Survival of civilization, Survival of humanity, Human survival, Survival of human, Survival of the human, Global collapse, Historical collapse, Catastrophic collapse, Global disaster, Existential threat, Catastrophic harm

We searched for these terms in the Wikipedia topic index and they resolved to the following terms (Table 2: Wikipedia Equivalent Terms), though many items did not resolve at all:

Table 2: Wikipedia Equivalent Terms

Global catastrophic risk*, Human extinction, Apocalypse, Doomsday, Doom, End of the world

* Wikipedia’s term “**Global catastrophic risk**” was a frequent destination for many Scopus search terms and can be seen as the Wikipedia equivalent to “Existential Risk”. See https://en.wikipedia.org/wiki/Global_catastrophic_risk

Wikipedia’s “Global catastrophic risk” page therefore seems to provide a good top-level page containing information about most existential risk content.

The number of Wikipedia equivalent terms is, however, relatively few and would provide a poor index to the breadth of topics falling within existential risk. At the bottom of the “Global catastrophic

risk" Wikipedia article, however, is a list of related topics which provides a more helpful index (Table 3: Wikipedia-related terms listed on “Global catastrophic risk”):

Table 3: Wikipedia-related terms listed on “Global catastrophic risk”

Future of the Earth, Ultimate fate of the universe, Doomsday Clock, Gray goo, Kinetic bombardment, Mutual assured destruction, Dead Hand, Doomsday device, Antimatter weapon, Synthetic intelligence / Artificial intelligence, Existential risk from artificial intelligence, AI takeover, Technological singularity, Transhumanism, Year 2000 problem, Malthusian catastrophe, New World Order (conspiracy theory), Nuclear holocaust, Nuclear winter, Nuclear famine, Cobalt bomb, Societal collapse, World War III, Climate change, Extinction risk from global warming, Tipping points in the climate system, Global terrestrial stilling, Global warming, Hypercane, Ice age, Ecocide, Human impact on the environment, Ozone depletion, Cascade effect, Supervolcano winter, Earth Overshoot Day, Overexploitation, Overpopulation, Human overpopulation, Extinction, Extinction event, Human extinction, Genetic erosion, Genetic pollution, Dysgenics, Pandemic, Biological agent, Transhumanism, Big Crunch, Big Rip, Coronal mass ejection, Gamma-ray burst, Impact event, Asteroid impact avoidance, Potentially hazardous object, Near-Earth supernova, Solar flare, Stellar collision, Eschatology, Buddhist eschatology, Hindu eschatology, Last Judgment, Christian eschatology, Book of Revelation, Islamic eschatology, Jewish eschatology, Ragnarok, Frashokereti, 2011 end times prediction, 2012 phenomenon, Apocalypse, Armageddon, Blood moon prophecy, End time, List of dates predicted for apocalyptic events, Nibiru cataclysm, Rapture, Revelation 12 sign prophecy, Third Temple, Alien invasion, Apocalyptic and post-apocalyptic fiction, List of apocalyptic and post-apocalyptic fiction, List of apocalyptic films, Climate fiction, Disaster films, List of disaster films, List of fictional doomsday devices, Zombie apocalypse

Each of these Wikipedia pages may provide a helpful visual topic marker to an existential risk topic that summarises the cluster the marker is located within (once processed and displayed on a concept map).

Before using all of these terms, however, it is necessary to apply the criteria for the existential risk bibliography, as set out in TERRA. These criteria require that all religious, mythological and fictional articles be removed from the bibliography. The goal of the concept map is to highlight those articles that ought to be in an existential risk bibliography so it's important to remove non-existential risk terms from the topic markers.

Removing non-existential-risk terms resulted in this reduced list (Table 4: Existential-risk-specific Wikipedia-related terms listed on “Global catastrophic risk”):

Table 4: Existential-risk-specific Wikipedia-related terms listed on “Global catastrophic risk”

Future of the Earth, Ultimate fate of the universe, Doomsday Clock, Gray goo, Kinetic bombardment, Mutual assured destruction, Dead Hand, Doomsday device, Antimatter weapon, Synthetic intelligence / Artificial intelligence, Existential risk from artificial intelligence, AI takeover, Technological singularity, Transhumanism, Year 2000 problem , Malthusian catastrophe, New World Order (conspiracy theory), Nuclear holocaust, Nuclear winter, Nuclear famine, Cobalt bomb, Societal collapse, World War III, Climate change, Extinction risk from

global warming, Tipping points in the climate system, Global terrestrial stilling, Global warming, Hypercane, Ice age, Ecocide, Human impact on the environment, Ozone depletion, Cascade effect, Supervolcano winter, Earth Overshoot Day, Overexploitation, Overpopulation, Human overpopulation, Extinction, Extinction event, Human extinction, Genetic erosion, Genetic pollution, Dysgenics, Pandemic, Biological agent, Transhumanism, Big Crunch, Big Rip, Coronal mass ejection, Gamma-ray burst, Impact event, Asteroid impact avoidance, Potentially hazardous object, Near-Earth supernova, Solar flare, Stellar collision, Apocalypse, Alien invasion

Adding in the “Wikipedia Equivalent Terms” (see Table 2, above) to this list, we end up with the following list of Wikipedia terms for use as topic markers (Table 5: Proposed topic markers):

Table 5: Proposed topic markers

Global catastrophic risk, Future of the Earth, Ultimate fate of the universe, Doomsday Clock, Gray goo, Kinetic bombardment, Mutual assured destruction, Dead Hand, Doomsday device, Antimatter weapon, Synthetic intelligence, Artificial intelligence , Existential risk from artificial intelligence, AI takeover, Technological singularity, Transhumanism, Year 2000 problem, Malthusian catastrophe, New World Order (conspiracy theory), Nuclear holocaust, Nuclear winter, Nuclear famine, Cobalt bomb, Societal collapse, World War III, Climate change, Extinction risk from global warming , Tipping points in the climate system, Global terrestrial stilling, Global warming, Hypercane, Ice age, Ecocide, Human impact on the environment, Ozone depletion, Cascade effect, Supervolcano , Volcanic winter, Earth Overshoot Day, Overexploitation, Overpopulation, Human overpopulation, Extinction, Extinction event, Human extinction, Genetic erosion, Genetic pollution, Dysgenics, Pandemic, Biological agent, Transhumanism, Big Crunch, Big Rip, Coronal mass ejection, Gamma-ray burst, Impact event, Asteroid impact avoidance, Potentially hazardous object, Near-Earth supernova, Solar flare, Stellar collision, Alien invasion, Apocalypse, Doomsday, Doom

After adding these terms and re-running the concept mapping, the majority of clusters had clear labels. However, several large existential-risk-related clusters seemed conspicuous by their lack of a label. Zooming-in revealed three clusters focused on the following topics: "Risk", "Insurance", and "Disaster". Appropriate Wikipedia pages for these three topics were located and the topics were added, creating a final list of topic markers (Table 6: Final topic markers):

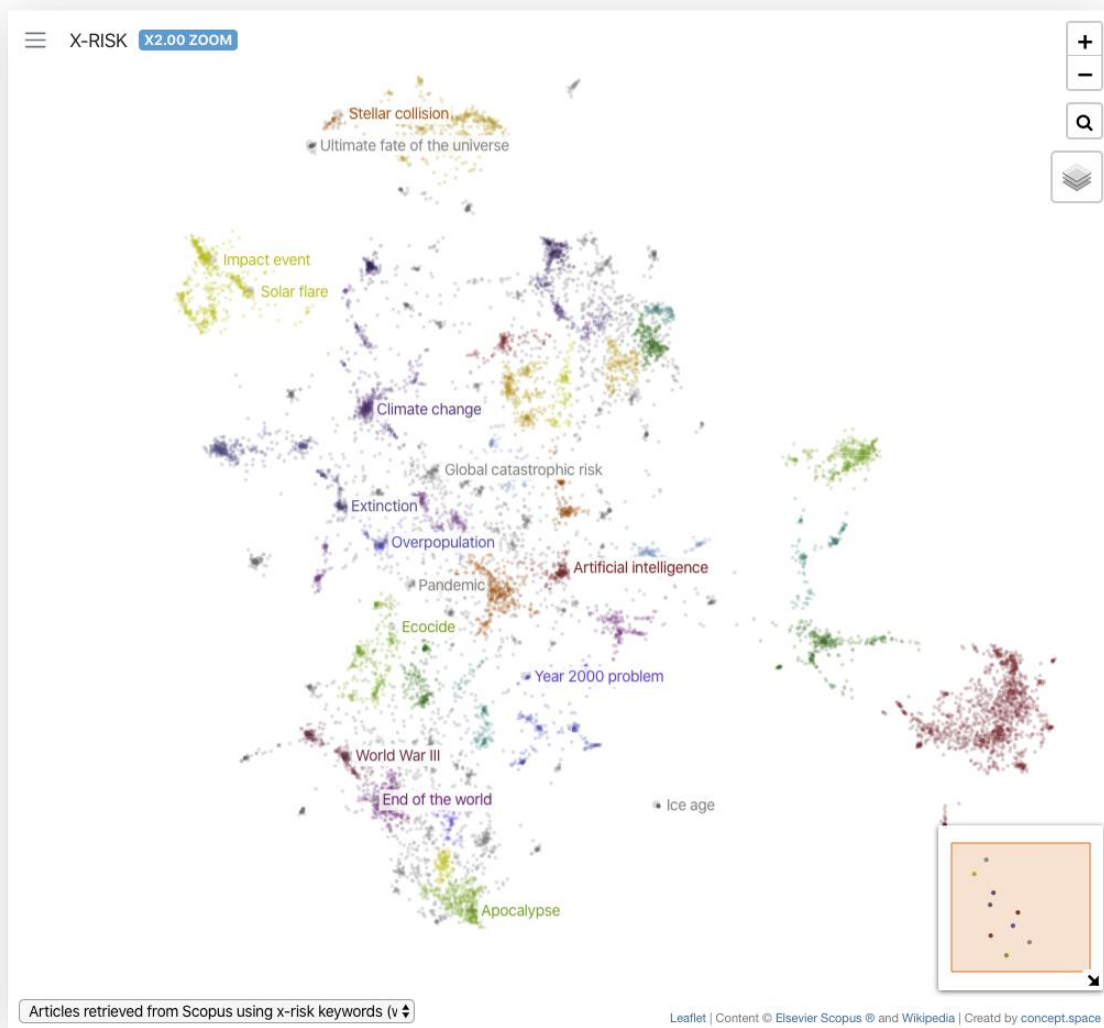
Table 6: Final topic markers

Global catastrophic risk, Future of the Earth, Ultimate fate of the universe, Doomsday Clock, Gray goo, Kinetic bombardment, Mutual assured destruction, Dead Hand, Doomsday device, Antimatter weapon, Synthetic intelligence, Artificial intelligence , Existential risk from artificial intelligence, AI takeover, Technological singularity, Transhumanism, Year 2000 problem, Malthusian catastrophe, New World Order (conspiracy theory), Nuclear holocaust, Nuclear winter, Nuclear famine, Cobalt bomb, Societal collapse, World War III, Climate change, Extinction risk from global warming, Tipping points in the climate system, Global terrestrial stilling, Global warming, Hypercane, Ice age, Ecocide, Human impact on the environment, Ozone depletion, Cascade effect, Supervolcano , Volcanic winter, Earth Overshoot Day, Overexploitation,

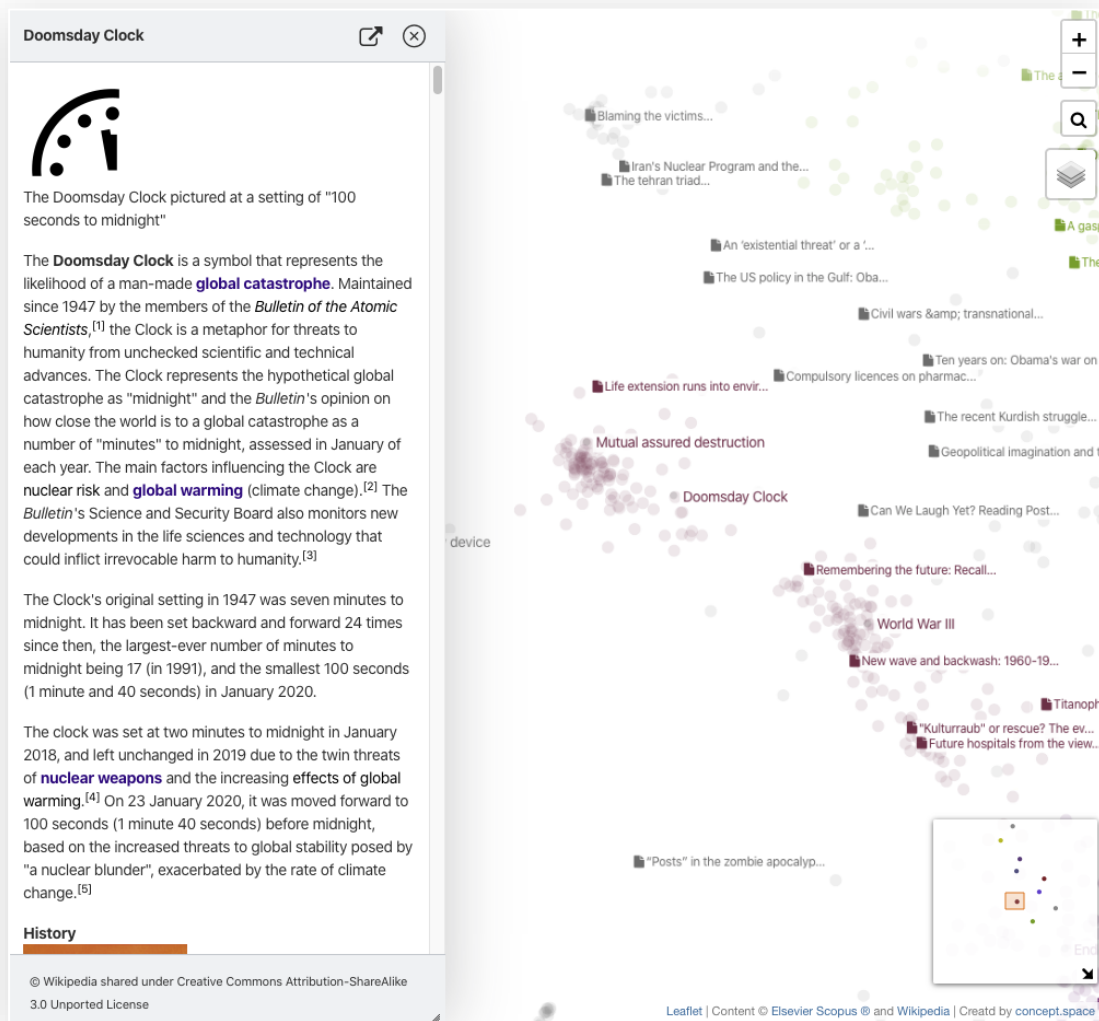
Overpopulation, Human overpopulation, Extinction, Extinction event, Human extinction, Genetic erosion, Genetic pollution, Dysgenics, Pandemic, Biological agent, Transhumanism, Big Crunch, Big Rip, Coronal mass ejection, Gamma-ray burst, Impact event, Asteroid impact avoidance, Potentially hazardous object, Near-Earth supernova, Solar flare, Stellar collision, Alien invasion, Apocalypse, Doomsday, Doom, End of the World, Risk, Insurance, Disaster


Running concept mapping again resulted in a second concept map **A2** with all Scopus-selected articles and relevant Wikipedia labels:

Concept Map A2



Clicking on a topic marker, eg. “Doomsday clock”, will reveal the Wikipedia page about that topic:



Clicking on the  symbol will open a new window displaying the original Wikipedia or DOI webpage for the topic/article (*note*: articles lacking a DOI link will not display this symbol).

The concept map appears to show a number of more or less well-clustered "concept islands" around certain topics. But as noted before, the process of dimensionality reduction that is involved in concept mapping may artificially force data points together, creating a false impression of reasonable clustering even if some data points have little in common. What is important to observe is the *relative* strength of the clustering.

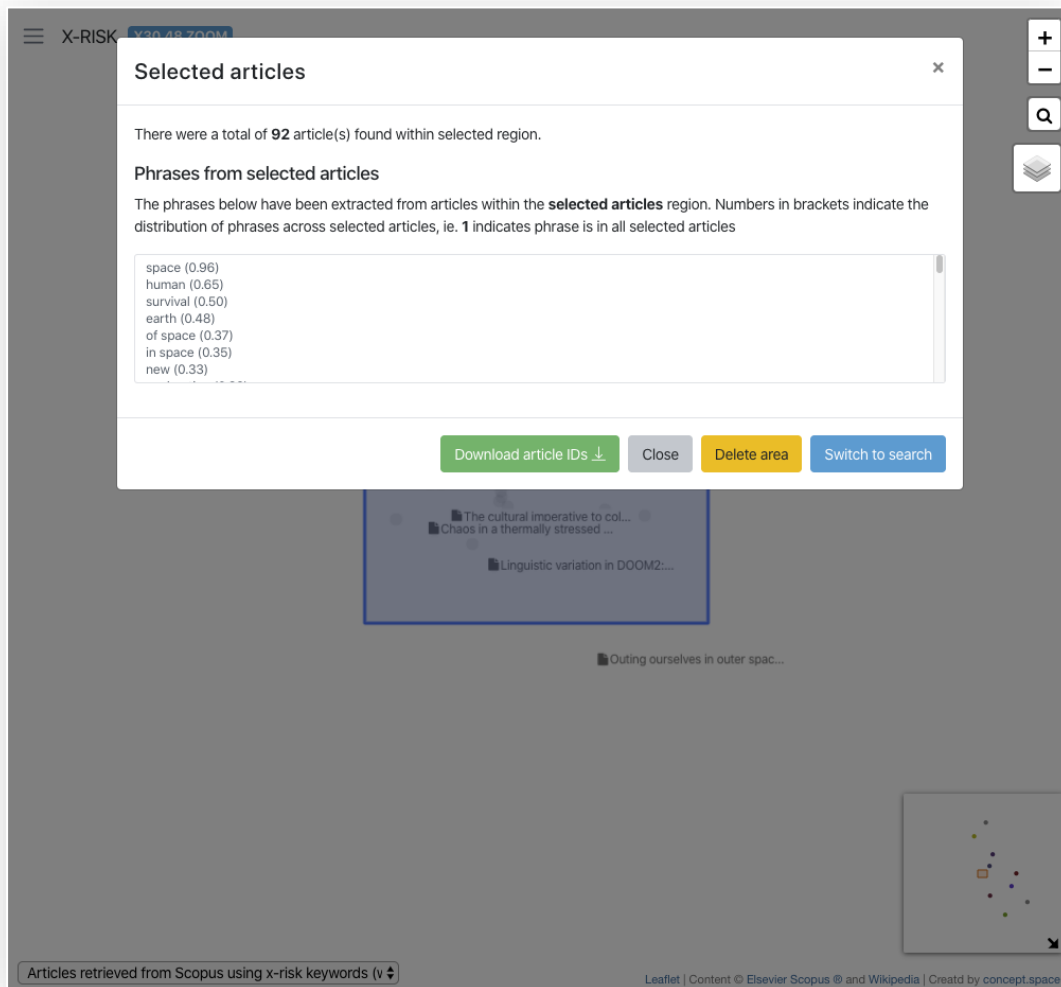
For example, if we zoom to x32, there is a strong clustering around "Sustainable space exploration":



Whereas the clustering is weaker around "Apocalypse" for the same magnification:

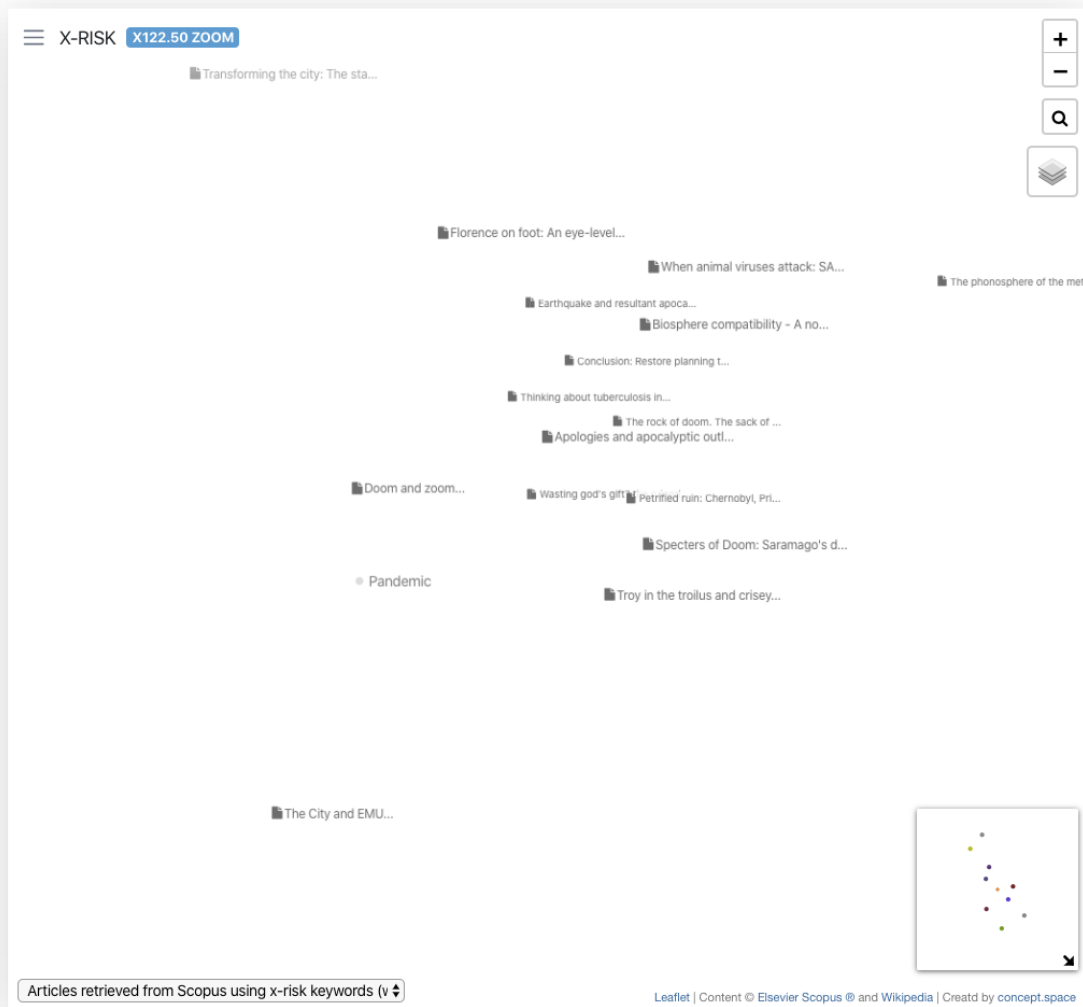


To evaluate a particular cluster, one can count the number of articles within a particular rectangular area by pressing **shift** on the keyboard and then **dragging a rectangle** with your mouse. The system will calculate the number of articles within the selected rectangle and also extract the most commonly-shared phrases within those articles (excluding Wikipedia articles):



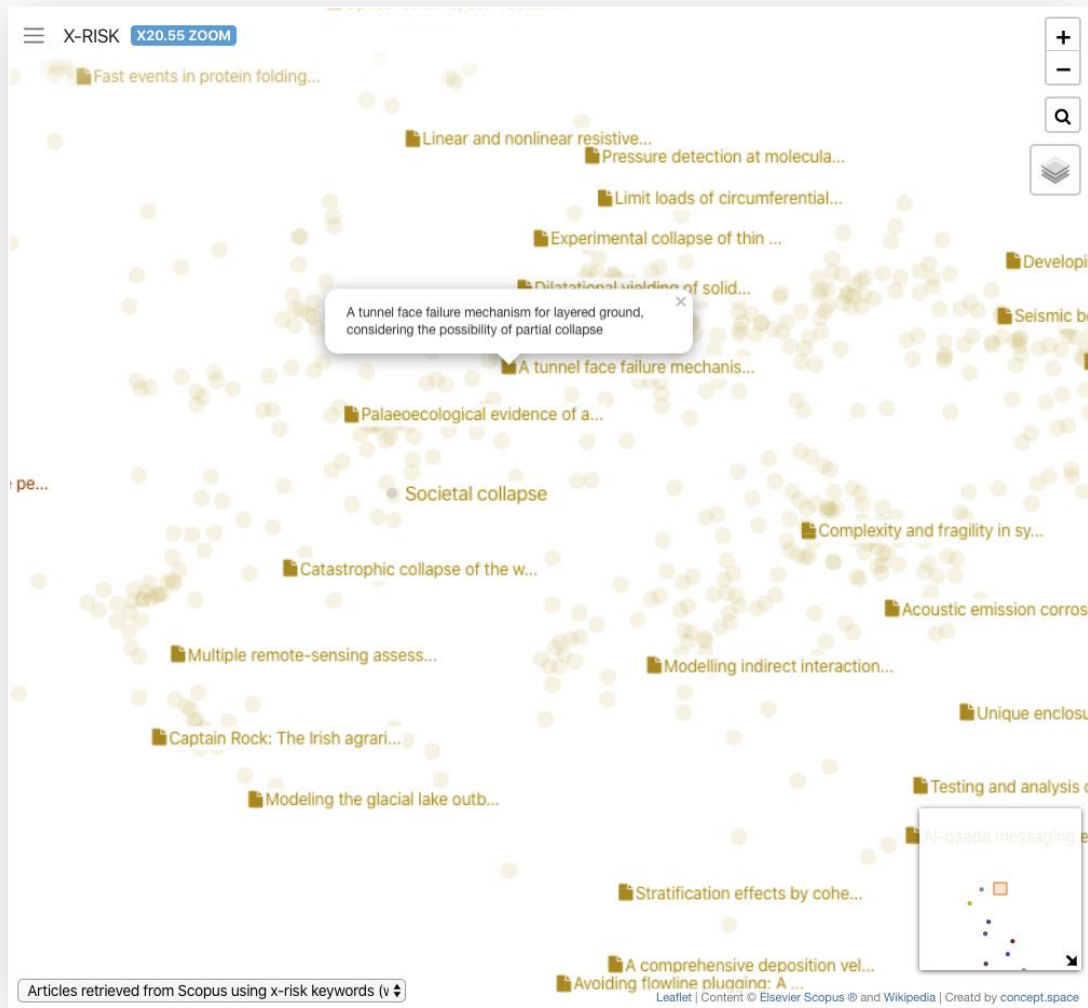
You will notice that flicking between concept maps **A1** and **A2** – which use the same identical text data except for the inclusion/exclusion of Wikipedia topic markers – presents a very different visual presentation of articles (though with broadly similar clustering). This is because adding Wikipedia content into the analysis affects how dimensionality reduction operates for *all* articles, which in turn affects where articles are positioned.

It should be noted that some clustering may represent a poor aggregation of articles. Take the clustering around “Pandemic” for example:



Very few of the articles near “Pandemic” are about infectious diseases and they seem mostly to do with disasters affecting cities. This may be partly a side-effect of dimensionality reduction which may force two very separate clusters in a multi-dimensional space to appear cheek-by-jowl next to each other. It may also be a consequence of using abstracts which are too short – short abstracts may lack sufficient textual information to accurately place an abstract in a concept space. It is suggested further research is carried out on different topic clusters to provide a metric for how accurately abstracts are located within their optimal topic cluster.

The concept mapping may also become confused by similar terms that have radically different meanings. If you zoom into “Societal *collapse*”, for example, there will be a number of articles about structural *collapse*:



These radical differences in meaning between similar-sounding phrases could be mitigated through the use of word-vector embeddings and neural networks, eg. Word2vec¹. However, for reasons of simplicity and computational time (Word2vec is computationally intensive), this approach was not pursued in this initial assessment of concept mapping.

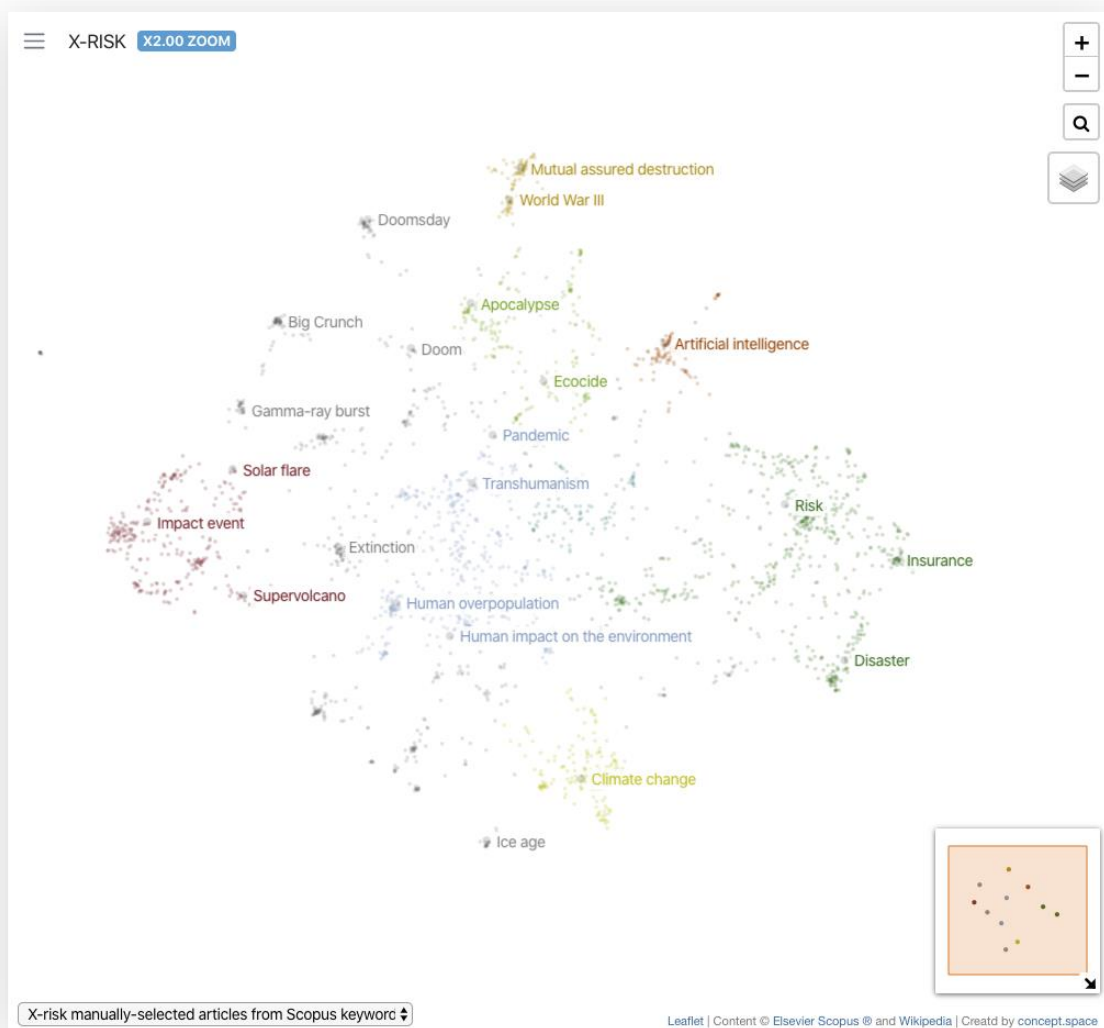
¹ <https://en.wikipedia.org/wiki/Word2vec>

4. Mapping manually-tagged XRisk terms

The wide range of phrases contained in the raw Scopus data feed may make it difficult to see texture within the **A1** and **A2** existential risk concept maps. This is because dense clusters of jargon-rich articles, eg. scientific papers on cancer, that may have nothing to do with existential risk may distort the dimensionality reduction process and flatten out the richness of existential-risk-specific phrases.

Several “**B**” concept maps were therefore created that looked solely at articles that had been manually tagged as "Existential Risk". The concept map **B1** displayed all articles manually tagged as "Existential Risk" by at least one person:

Concept Map B1

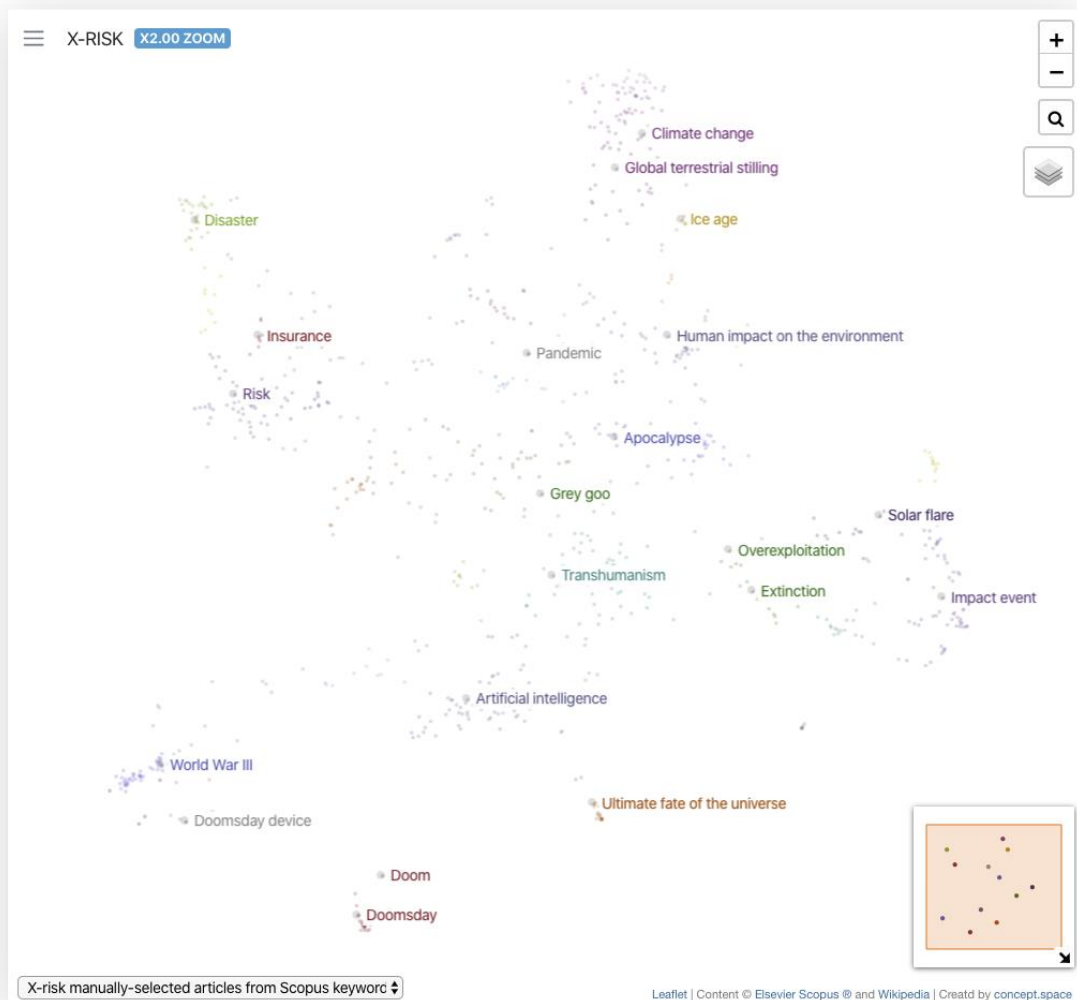


This map does, however, include articles that have been classified as existential risk by a single reviewer. During the initial exploration of tagged Scopus results, it became clear several articles had

been incorrectly classified as “Existential Risk” by single reviewers. These incorrectly tagged articles may skew the conclusions drawn from the map. One may conclude, for example, that a cluster of articles should not be excluded from the existential risk bibliography on account of the presence of one (incorrectly-tagged) article within the cluster. The concept map **B1** makes no distinction between articles that are classified as existential risk by a single reviewer, ie. least confident, or by multiple reviewers, ie. more confident.

It was therefore decided to create a second concept map **B2** that only considered articles that have been tagged as existential risk by more than one reviewer:

Concept Map B2



It is clear from this map there are some seemingly well-defined clusters among certain subject areas, eg. *climate change*, *risk/insurance*, *World War III*, *doom/doomsday*, *ultimate fate of the Universe*, *solar flare/impact event*, *overexploitation/extinction*, *artificial intelligence*. While this may be due to the larger volume of articles written about these topics, there appear to be clear differences of density between different topic areas, eg. *doom/doomsday* and *World War III* are more densely clustered than

transhumanism. More research is needed to ensure the denser clusters represent genuine clusters of articles around a specific topic; increasing the number of manually tagged existential risk articles in general would certainly help in this regard.

Assuming the visible clusters do indeed represent genuine topic-specific clusters of articles, the clustering that seems evident suggests one way in which the TERRA search strategy could be improved – through the creation of cluster-specific search strategies that deepen the number of articles in each cluster. This could be achieved as follows:

- **Keyphrase searching:** Use the existing clusters to help define a range of keywords for each topic cluster that more accurately select general Scopus articles for that cluster. This has the advantage of being easily reproducible, a key requirement of TERRA’s vision.
- **Machine Learning (ML) classification:** Use the existing clusters to help define training sets for each topic that would then be used to create topic-specific Machine Learning classifiers. An example of how this can be done can be seen at <http://crab3.lionproject.net²>, which contains roughly 50 topic-specific ML classifiers.

In both cases, the advantage of a cluster-specific search strategy is that additional existential risk articles might be retrieved that are not captured by the existing Scopus search criteria.

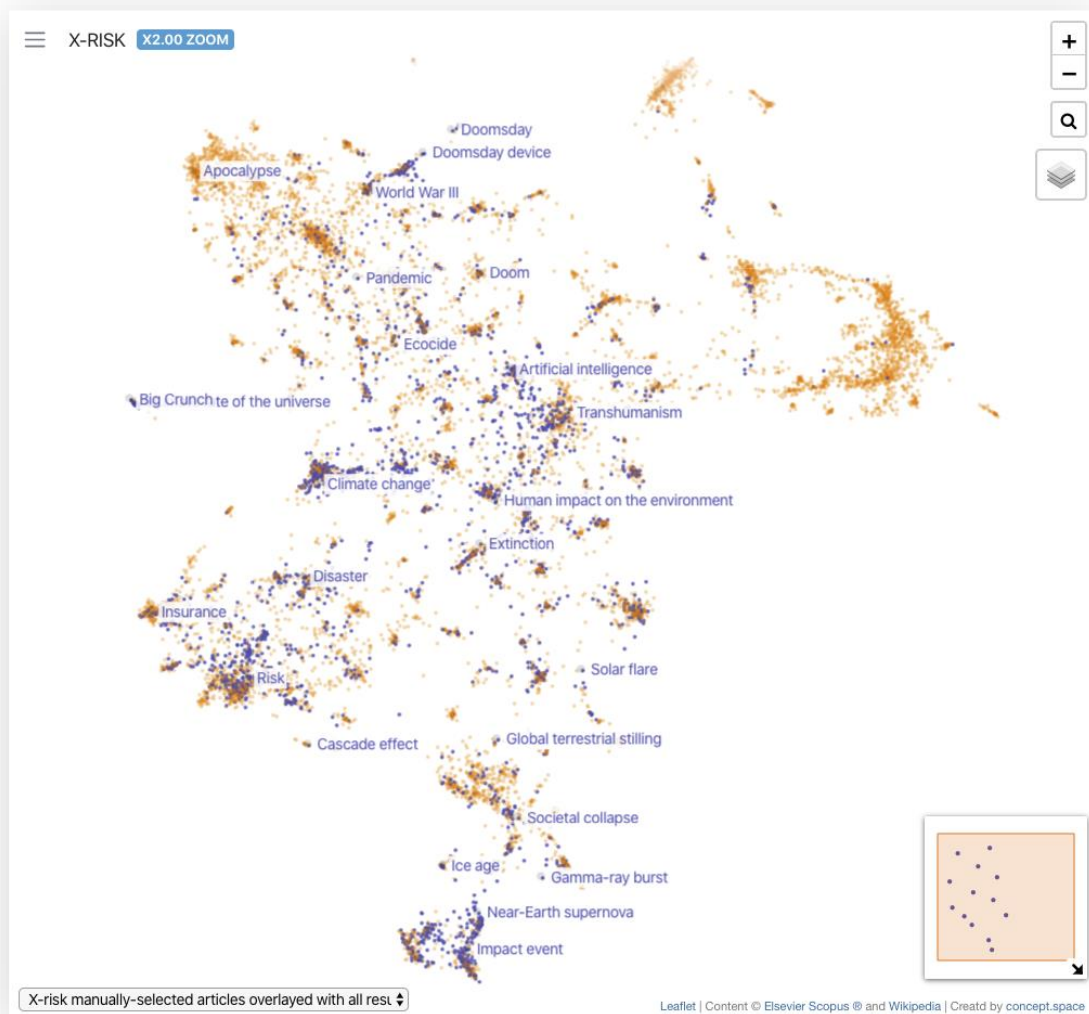
² Version 3 release of CRAB was developed by the author

5. Comparing tagged and untagged articles

A major motivation for carrying out concept mapping on TERRA text content was to compare how a concept map changes from “All Scopus-retrieved articles” (“A” maps) to “All manually-tagged existential risk articles” (“B” maps). However, “A” concept maps are so radically different to “B” concept maps it is difficult to draw substantive conclusions. It was therefore decided to create a bespoke visualization that identified manually tagged existential risk articles within an overall context of all Scopus-retrieved articles. These maps have the naming convention of **A[*],B[*]**, for example **A+,B2**.

The first such comparison map **A,B1** highlighted *all* manually-tagged existential risk articles (regardless of number of reviewers) on the map of all Scopus-retrieved articles:

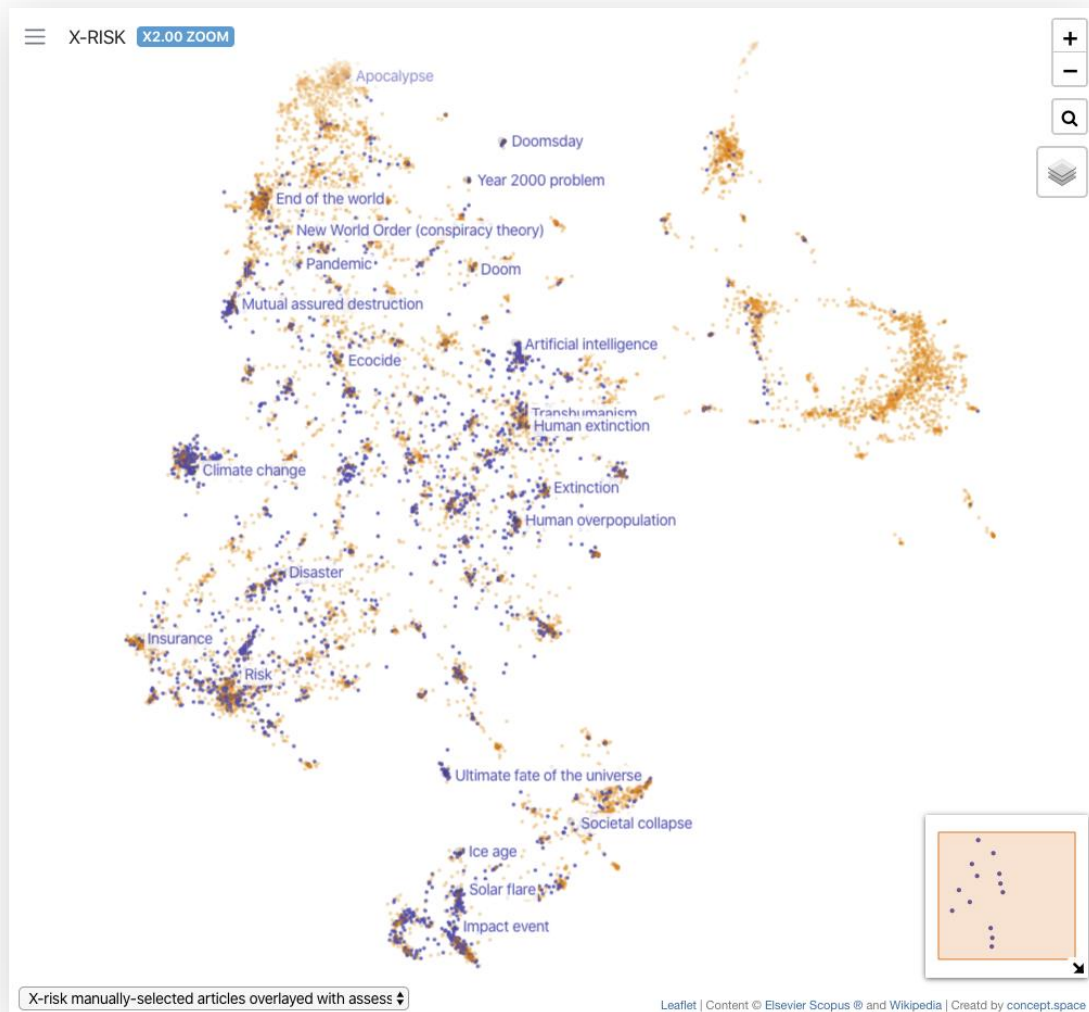
Concept Map A,B1



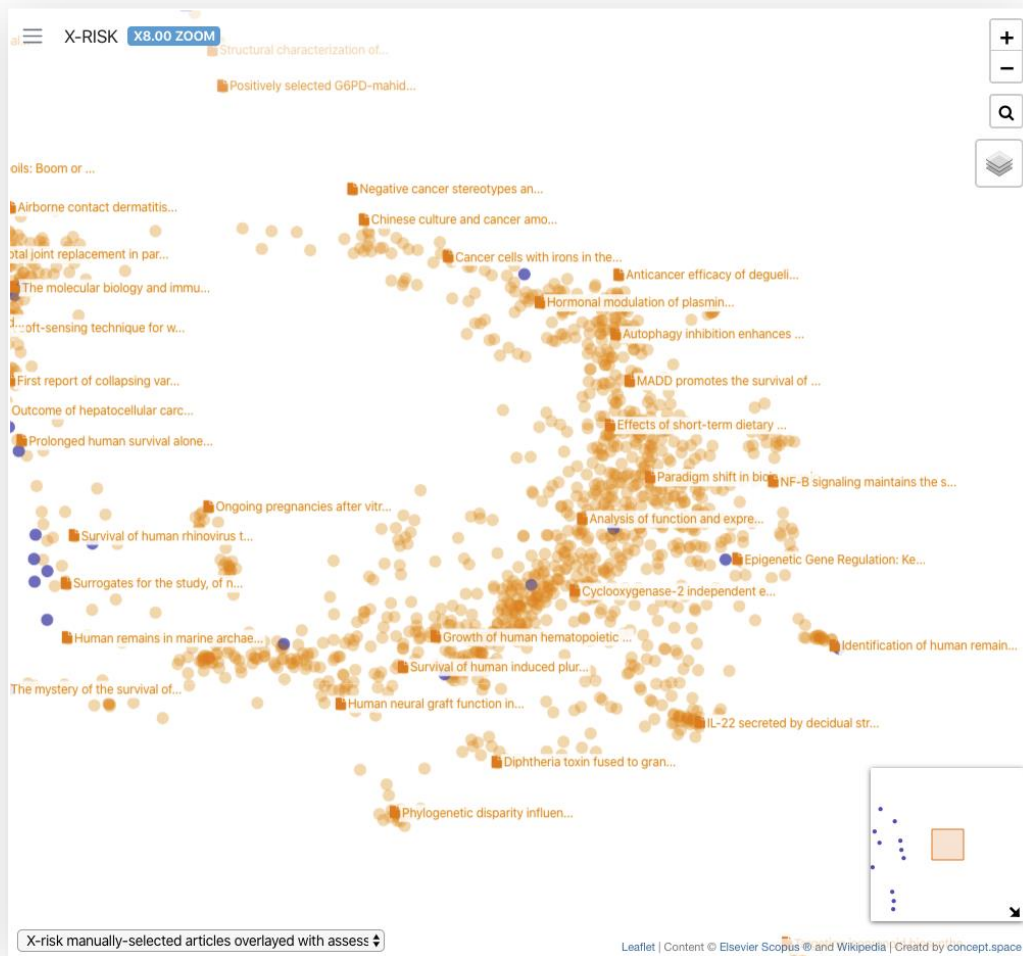
The blue dots indicate those articles that have been manually tagged as existential risk while the orange dots indicate all other articles, either not tagged as existential risk or not assessed by any user.

To reduce the number of orange dots, a second map was created that compared those articles that had been assessed as existential risk by at least one reviewer with those articles that had been assessed and had been classified as *not* being existential risk. This map is termed **A+,B1** (where **A+** indicates the reduced context of “Assessed Scopus-retrieved articles only”):

Concept Map A+,B1



You will notice that towards the top-right of the screen is a large area of orange with relatively few blue dots. On closer examination, these articles are about general cell biology, medical research and cancer.



It is worth noting that in the PubMed database of medical research, around 16% of all articles are about cancer (see <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5345838/>).

Zooming in closer, some of the blue-coloured articles do not appear to be *clearly* about existential risk. For example:

Enhanced Stem Cell Differentiation and Immunopurification of Genome Engineered Human Retinal Ganglion Cells

Zack D.J., Welsbie D.S., Cheng J., Berlinicke C.A., Mitchell K.L., Sluch V.M., Liu M.M., Chamling X.
Stem Cells Translational Medicine
2017

© 2017 The Authors Stem Cells Translational Medicine published by Wiley Periodicals, Inc. on behalf of AlphaMed Press Human pluripotent stem cells have the potential to promote biological studies and accelerate drug discovery efforts by making possible direct experimentation on a variety of human cell types of interest. However, stem cell cultures are generally heterogeneous and efficient differentiation and purification protocols are often lacking. Here, we describe the generation of clustered regularly-interspaced short palindromic repeats (CRISPR)-Cas9 engineered reporter knock-in embryonic stem cell lines in which tdTomato and a unique cell-surface protein, THY1.2, are expressed under the control of the retinal ganglion cell (RGC)-enriched gene BRN3B. Using these reporter cell lines, we greatly improved adherent stem cell differentiation to the RGC lineage by optimizing a novel combination of small molecules and established an anti-THY1.2-based protocol that allows for large-scale RGC immunopurification. RNA-sequencing confirmed the similarity of the stem cell-derived RGCs to their endogenous human counterparts. Additionally, we developed an in vitro axonal injury model suitable for studying signaling pathways and mechanisms of human RGC cell death and for high-throughput screening for neuroprotective compounds. Using this system in combination with RNAi-based knockdown, we show that knockdown of dual leucine kinase (DLK) promotes survival of human RGCs, expanding to the human system prior reports that DLK inhibition is neuroprotective for murine RGCs. These improvements will facilitate the development and use of large-scale experimental paradigms that require numbers of pure RGCs that were not previously

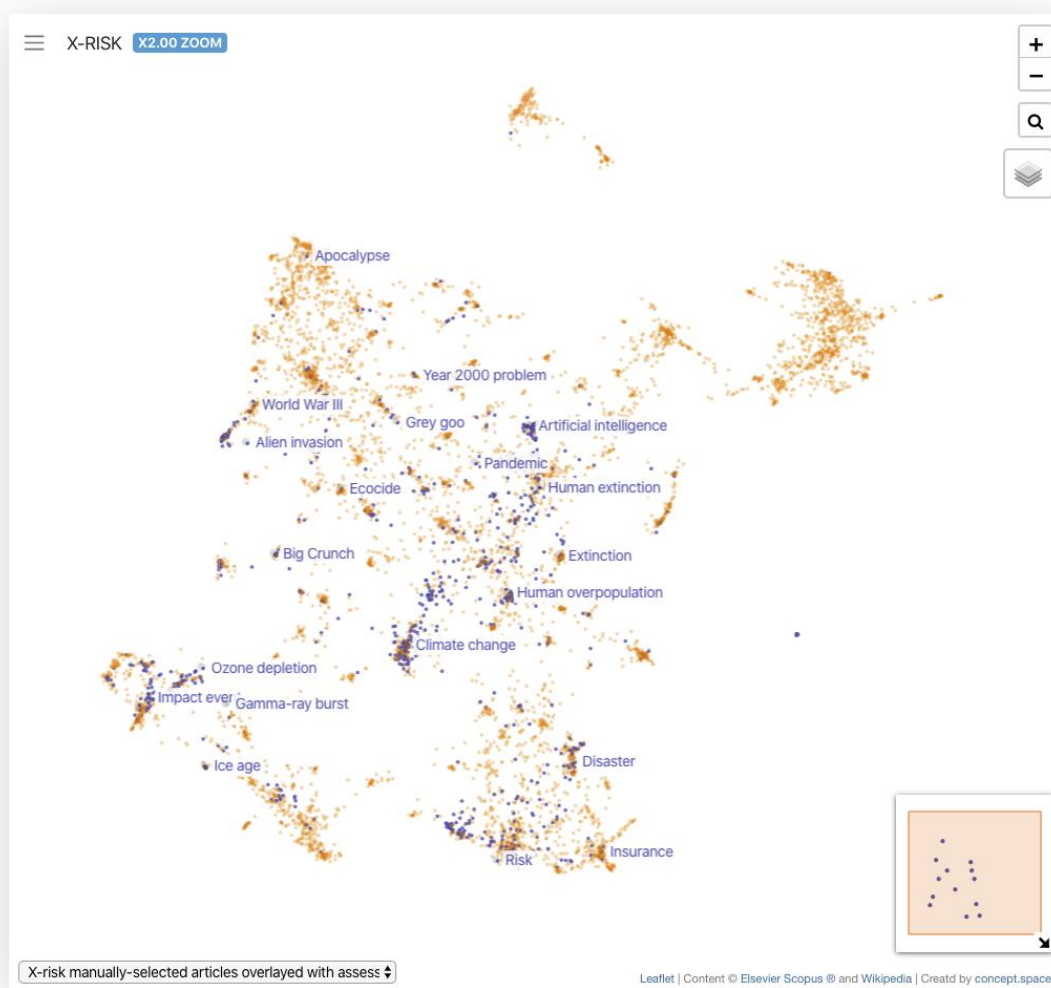
© Elsevier Scopus®

Leaflet | Content © Elsevier Scopus® and Wikipedia | Created by concept.space

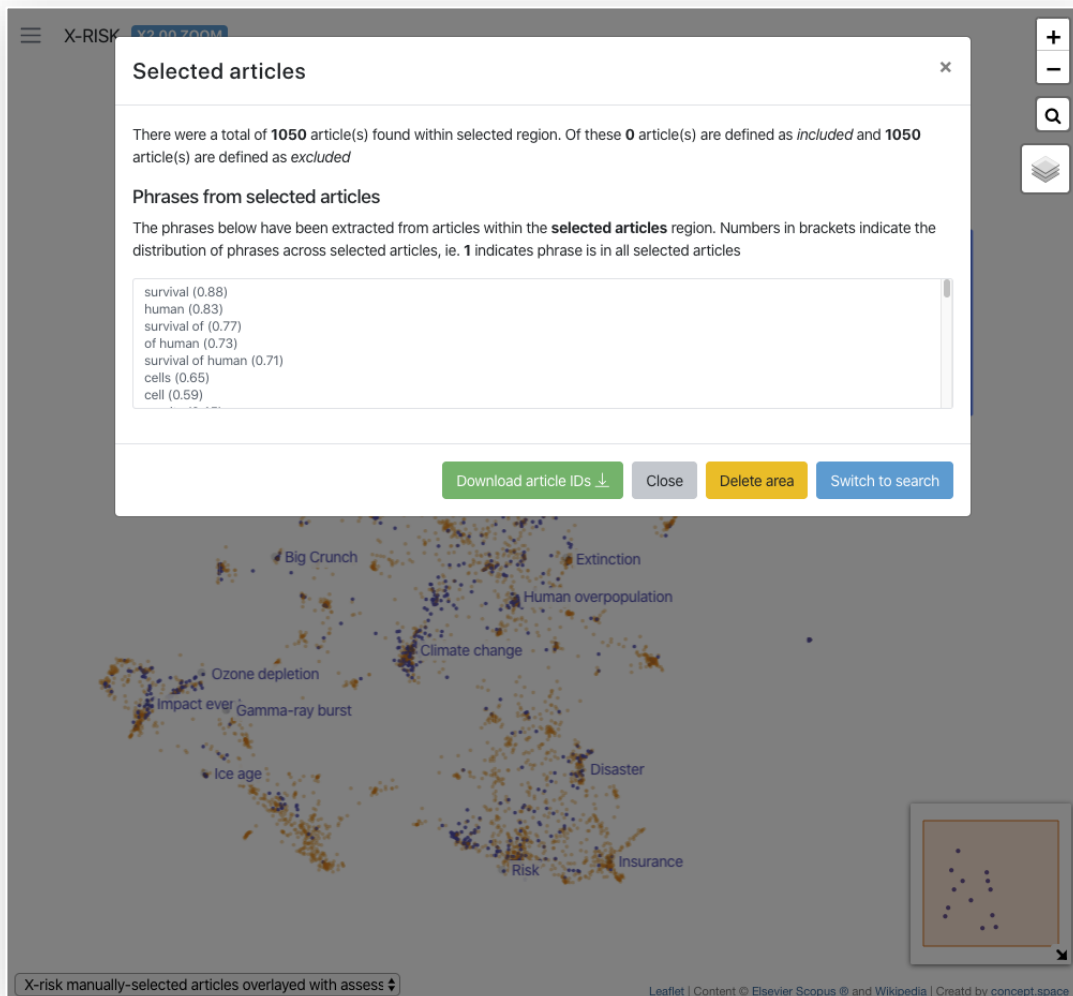
This appears to be a potential example of an incorrect single-reviewer assessment skewing the conclusions that might be drawn from the concept map - the orange cluster should not be removed, it might be argued, because it contains existential risk articles. Removing the incorrect assessments might lead to a different conclusion, ie. that the cluster should be excluded.

A further concept map **A+**, **B2** was therefore created that looked only at results where at least two reviewers had made an assessment:

Concept Map A+,B2



The large orange cluster towards the top-right of the map now seems free of any existential risk articles. This can be confirmed by shift-dragging a rectangle over the area:



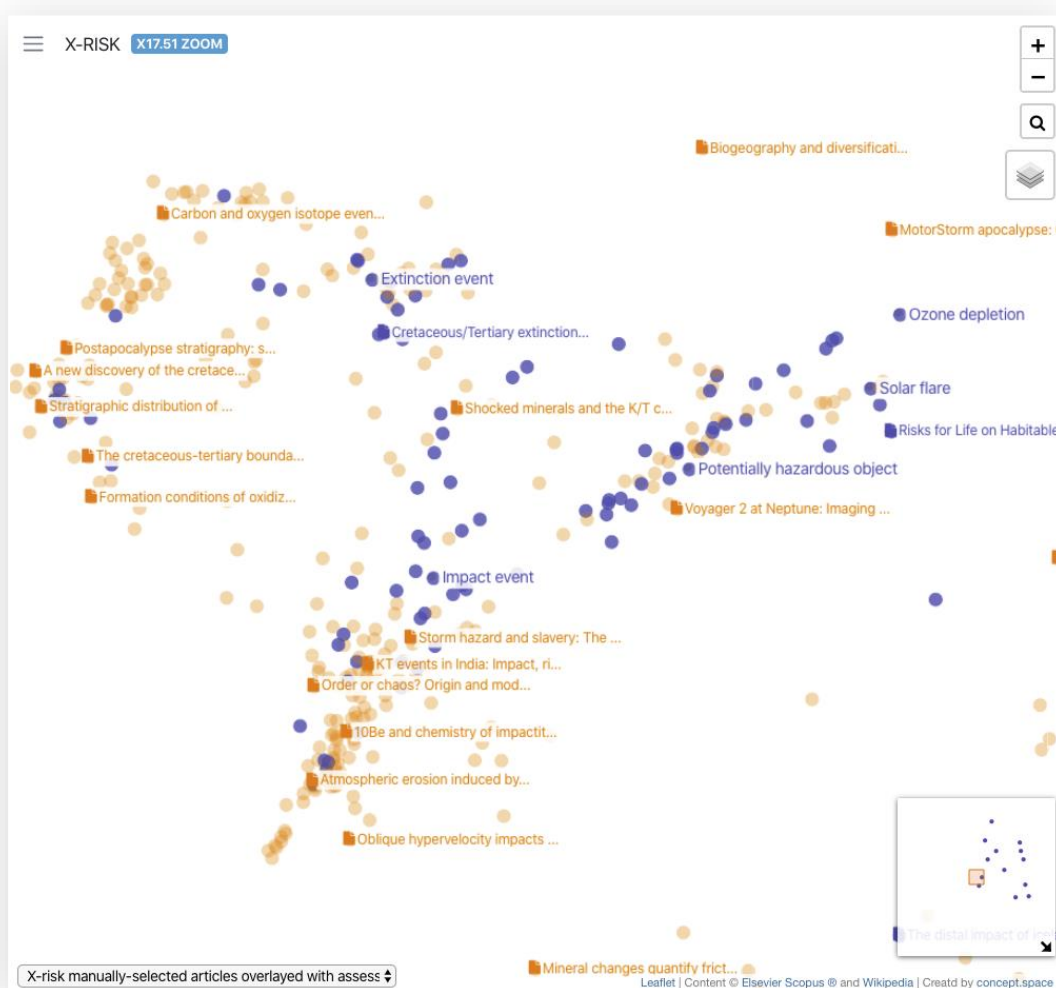
Note the text that is displayed: “*There were a total of 1050 article(s) found within selected region. Of these 0 article(s) are defined as included and 1050 article(s) are defined as excluded*”. An “included” article is one that has been manually classified as existential risk while an “excluded” article is one that has not - either because it has not been assessed at all or because it has been assessed and found not to be an existential risk article. We shall refer to this particular problem cluster towards the top-right of map **A+,B2** as “Problem Cluster A”.

The existence of such problem clusters suggests a further strategy to improve the TERRA bibliography – specifically excluding clusters that may have crept into the Scopus results through the use of an ambiguous search phrase. In the screengrab above, for example, you can see the phrase “Survival of human” which is one of the original Scopus search terms in **Table 1: Scopus Search Terms**. It is, however, a term frequently used in medical research, for example when talking about the “*survival of human cells*”, the “*survival of human organs*”, the “*survival of human cancer cells*”, etc.

There are two ways such a cluster might be excluded, which follow the same approach one might use when trying to *include* a specific topic cluster:

- **Keyphrase searching:** Define a range of keywords that accurately capture the majority of articles in the cluster, while avoiding capturing articles outside that cluster. This has the advantage of being easily reproducible, a key requirement of TERRA’s vision.
- **Machine Learning (ML) classification:** Use the existing cluster to define a Machine Learning training set for that cluster which is then used to create a cluster-specific Machine Learning classifier.

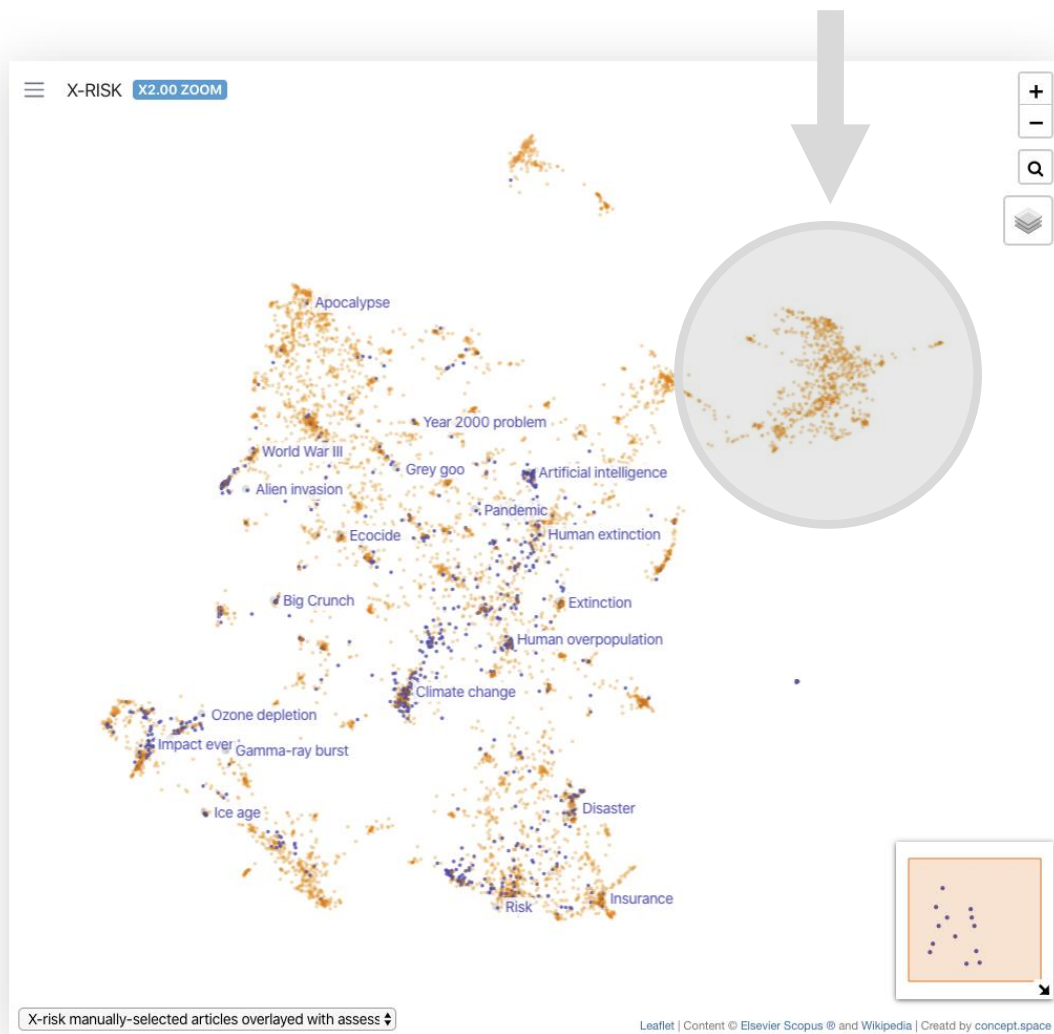
In both cases, the use of techniques to exclude “rogue” clusters – large clusters of articles that trivially have nothing to do with existential risk - would help focus reviewers on more “challenging” areas of classification. In these challenging areas existential risk articles are well mixed in with non-existential risk articles on the concept map, suggesting a similarity of key phrases in both types of article that requires more sophisticated human interpretation to delineate xrisk versus non-xrisk. For example, around “Impact event”:



An attempt to exclude problem clusters using keyphrase searching will now be described.

6. Excluding problem clusters

“Problem Cluster A” to the top-right of concept map A+,B2 will be used to test how we might remove problem clusters using keyphrase searching:



If we shift-drag a rectangle over this cluster, we can retrieve the most commonly shared keyphrases in the selected articles:

Selected articles

There were a total of **873** article(s) found within selected region. Of these **0** article(s) are defined as *included* and **873** article(s) are defined as *excluded*

Phrases from selected articles

The phrases below have been extracted from articles within the **selected articles** region. Numbers in brackets indicate the distribution of phrases across selected articles, ie. **1** indicates phrase is in all selected articles

- survival (0.93)
- human (0.92)
- survival of (0.88)
- of human (0.84)
- survival of human (0.82)
- cells (0.76)
- cell (0.68)

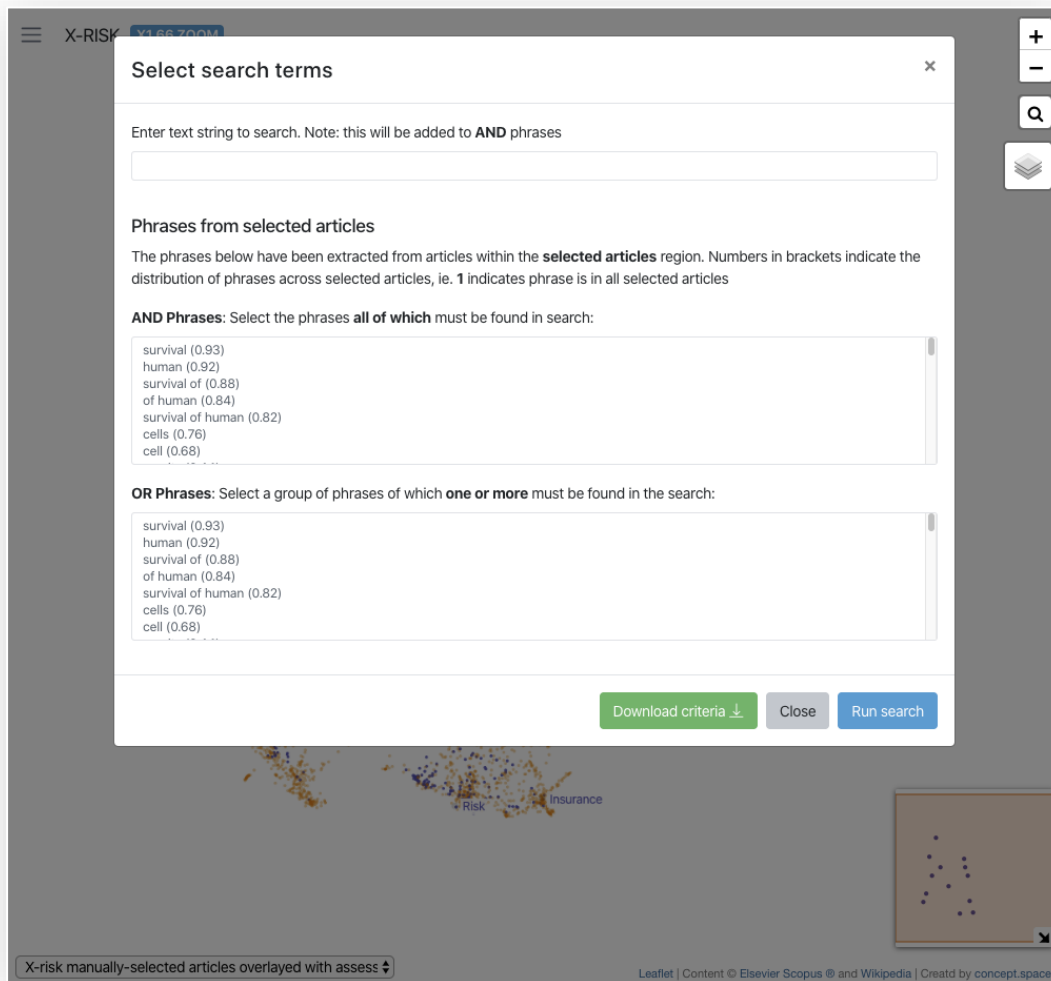
Download article IDs ↴ Close Delete area Switch to search

Big Crunch Extinction Human overpopulation
 Ozone depletion Climate change
 Impact event Gamma-ray burst Disaster
 Ice age Risk Insurance

X-risk manually-selected articles overlaid with assess

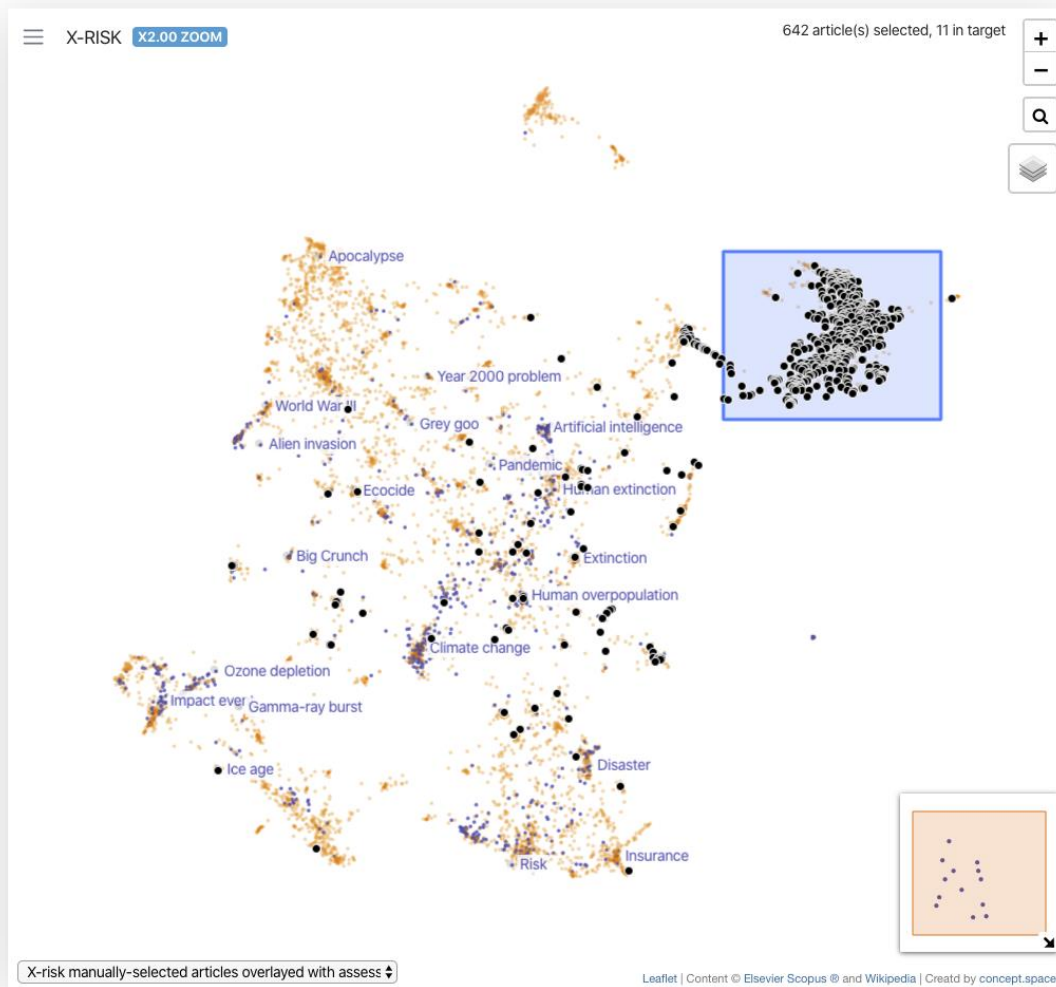
Leaflet | Content © Elsevier Scopus ® and Wikipedia | Created by concept.space

If we then click on “Switch to search”, we can build a complex query that can be used to select many of the articles within the problem cluster:



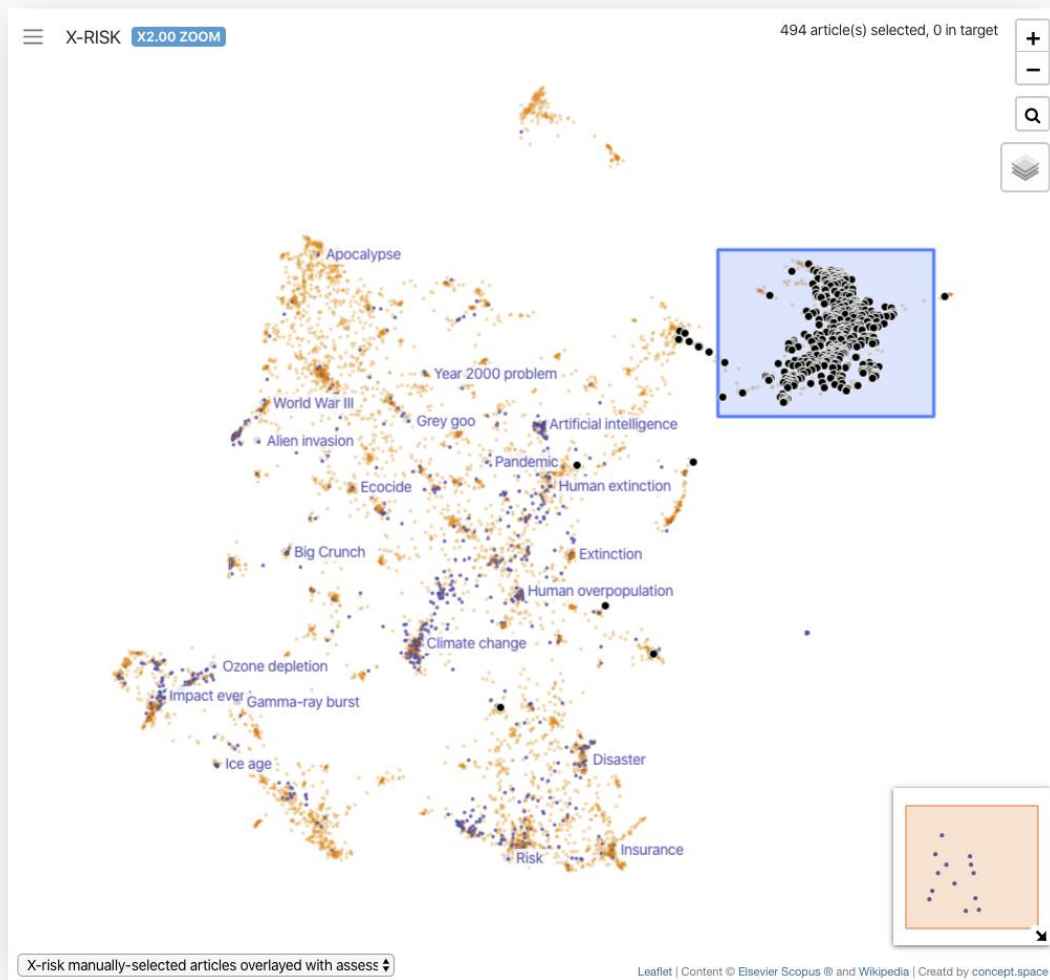
The list of “AND” phrases are phrases **all of which** must be found in a prospective search, whilst “OR” phrases are phrases of which **one or more** must be found.

As mentioned before, “*survival of human*” ranks highly as a commonly shared phrase in “Problem Cluster A”. It therefore makes a good choice as a mandatory “AND” term that will capture most of the articles in the selected rectangle. If you select *survival of human* under “AND Phrases” and click “Run search”, you will see the articles which are selected based on this criteria:

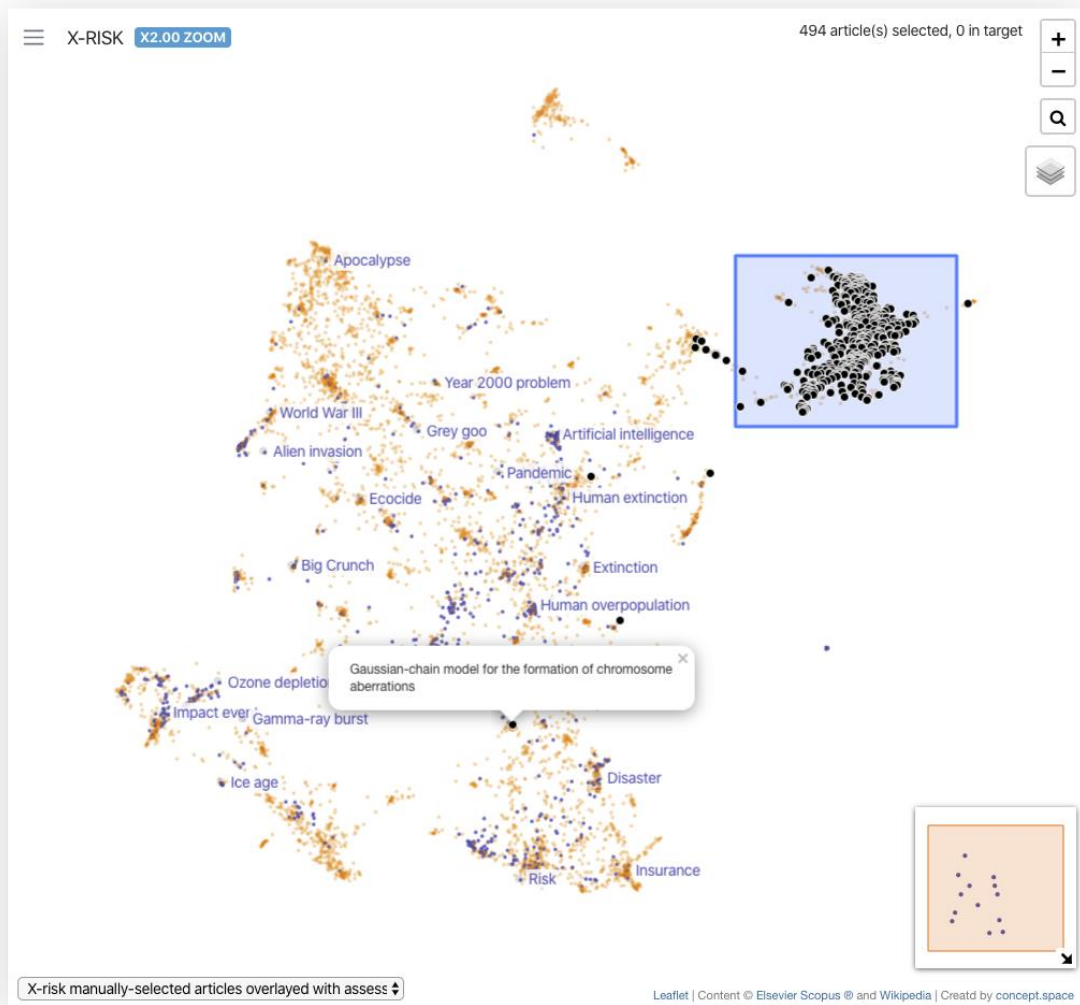


Many of the articles inside the rectangle have been selected, but also many articles outside of it as well. Given our goal was to remove “Problem Cluster A” only, it’s important to restrict articles as much as possible to those inside the rectangle.

To further restrict the article selection, you can add one or more “OR Phrases”. Given the medical nature of many of the articles inside the rectangle, the following medical-specific phrases were selected under “OR Phrases”: *cells, cancer, apoptosis, in vitro*. This significantly reduced the spread of selected articles:



There are still some articles outside the rectangle but you can check whether or not they are existential risk articles by floating your mouse over each black (selected article) circle:



You can also see the number of “target”, ie. included, articles among the total number of selected articles (represented by the black circles) in the top-right of the screen (“494 article(s) selected, 0 in target”).

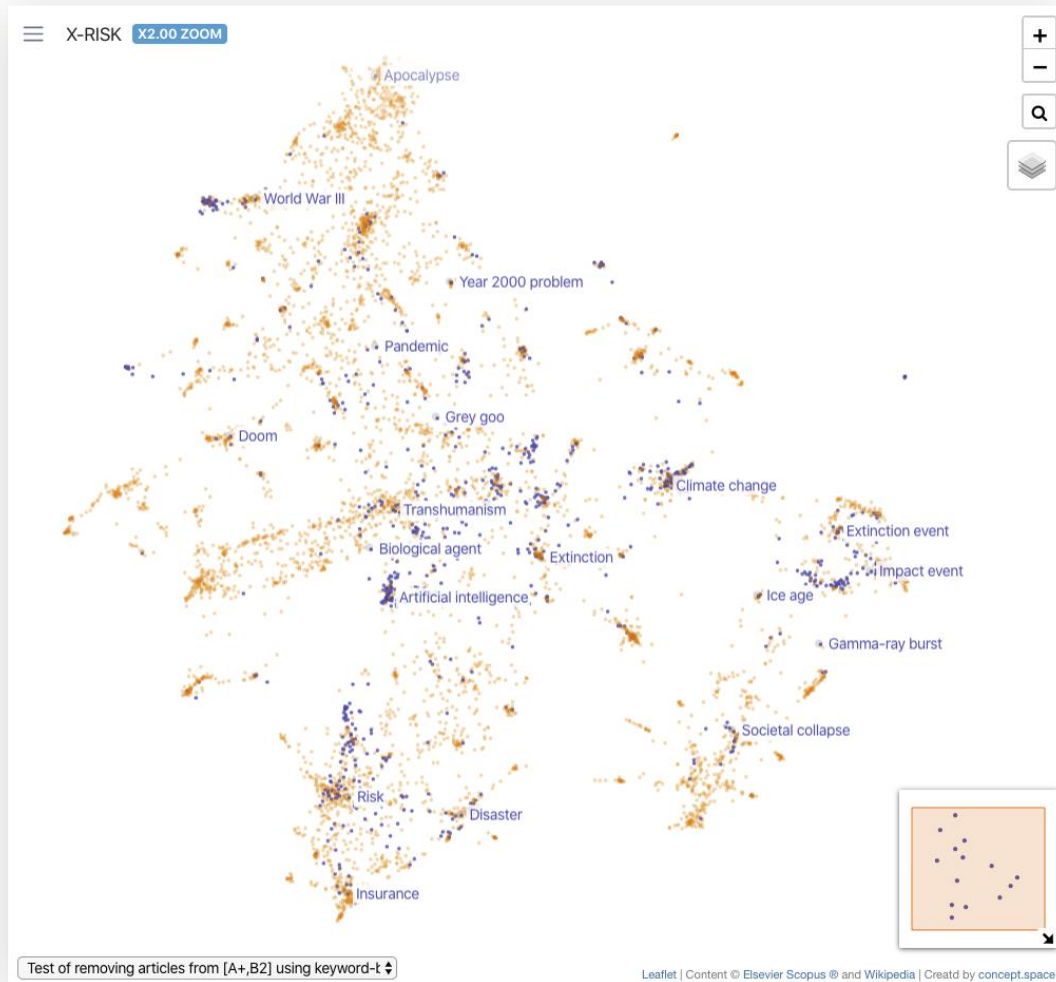
With the search criteria in reasonable shape, we can now download the search criteria file. Click on the magnifying glass icon (“Search and highlight”) and then click “Download criteria”. This will download a “searchphrases” JSON file with the search criteria as an array:

```
{
  "and": [
    "survival of human"
  ],
  "or": [
    "cells",
    "apoptosis",
    "cancer",
    "in vitro"
  ]
}
```

These “searchphrases” files can be used to filter out articles that exist in problem clusters. While not currently implemented in TERRA, integration of “searchphrases” exclusion files within the existing TERRA data flow should be a straightforward function to implement.

Using the searchphrases file generated above, it was possible to filter out many of the articles from “Problem Cluster A” in order to create concept map **R1**:

Concept Map R1



“Problem Cluster A” appears to have disappeared – at least as a conspicuously large cluster - although there are now smaller problem clusters to the centre left of the map. It is now potentially possible to repeat the cluster exclusion process with further keyphrase selection in order to remove further irrelevant clusters.

Beyond removing “Problem Cluster A”, however, it is not clear whether subsequent cluster exclusions will be anywhere near as radical, in terms of removing irrelevant articles. Nevertheless, by reducing the weight of non-existential-risk technical terms we give more space for existential terms to come to the fore – and the concept maps that result may highlight further useful keyphrase optimizations.

Conclusions

The concept mapping carried out on TERRA articles suggests a highly textured space of existential risk concepts, with articles on some existential risk topics being densely clustered while others are more diffusely spread.

The concept mapping also provides a useful visual indication of some of the flaws of the existing search strategy, for example the inclusion of ambiguous terms and the possible existence of incorrect single-reviewer assessments. It suggests possible ways in which the search strategy could be improved by focusing on cluster-specific inclusion and exclusion techniques. Cluster-specific selection may be carried out with an intuitively straightforward keyphrase selection approach or a more advanced machine learning approach.

It is proposed that a cluster- or topic-specific approach to manual reviewing – whereby articles within recognisable existential risk clusters are prioritised above articles within potentially “rogue clusters” - may avoid the risk of presenting manual reviewers with trivial cases of non-existential-risk articles. At the very least there is scope for creating two separate streams of manual reviewing, where one stream involves entry-level reviewers classifying trivially non-existential-risk articles, while a second stream allows advanced users to decide upon more challenging classifications.

In terms of promoting reviewer engagement, concept mapping also offers a useful “gamification” dimension to the reviewing process. In the current version of TERRA, users are presented with a linear progression of articles to review and the process of making progress through this list may appear daunting or uninspiring. A concept mapping approach, by contrast, offers the possibility of reviewers “colouring in” particular chunks of the map – a chunk may represent a convenient amount of reviewing to do in one sitting. It is also possible to create a literal game of the space of existential risk concepts and articles in which users interact with concepts (for example, see <https://vimeo.com/327236733>)

Appendix 1 - Description of concept mapping process

1. Text tokenization

During the text tokenization stage, the text content of all Scopus-searched articles and Wikipedia topics is converted into numbers, one number for each text phrase (or “token”):

https://en.wikipedia.org/wiki/Lexical_analysis#Tokenization

The cumulative set of numbers for each article abstract or topic represents a “Bag of Words”, where the count of each text token in each article represents the value on that text token’s axis:

https://en.wikipedia.org/wiki/Bag-of-words_model

The complete set of axes for all articles represents a multi-dimensional space, with every article represented by a point in that space. Articles with more text tokens in common will be “closer” to each other than other articles with fewer text tokens in common.

2. Apply dimensionality reduction

The process of “dimensionality reduction” reduces the multidimensional space of all text articles down to a few dimensions – in our case two:

https://en.wikipedia.org/wiki/Dimensionality_reduction

A number of different dimensionality reduction algorithms exist including PCA, t-SNE and UMAP. In this project, we used UMAP:

<https://umap-learn.readthedocs.io/en/latest/>

3. Apply clustering to colour-code similar clusters

After dimensionality reduction has completed, a clustering algorithm is used to colour-code points in the same cluster:

https://en.wikipedia.org/wiki/Cluster_analysis

A number of different clustering algorithms exist. In this project, we used HDBSCAN:

<https://github.com/scikit-learn-contrib/hdbscan>