

Solution and Recommendations

From Digital Academy Team 9 – The Avengers

By: Bo Peng

June 3rd, 2019

|  |  |  |
| --- | --- | --- |
| **Writer name** | **Update date** | **Contents changes** |
| Bo Peng | May 30, 2019 | Created document with notes for problem, background, recommendations, prototype, team, stakeholder, and interviewees. - draft |
| Bo Peng | June 3, 2019 | Refined and added more details in most sections, added architecture diagram by Marwan, finished version 1 |
| Bo Peng | June 5, 2019 | Updated contents, added dashboard graphics, References, Glossary finished version 2 |
| Bo Peng | June 6, 2019 | Updated contents, added /updated web interface images, Tesseract by Tim, finished version 2.1 |
| Bo Peng | June 7, 2019 | Added detail architecture diagram by Marwan, and added the section describe the detail architecture, code repository URL for future developers. |
| Bo Peng | June 10, 2019 | Edited the document in general; added section of technologies used. |
| Bo Peng | June 11, 2019 | Added screen shots from Julien. |
|  |  |  |

Document history

I. Problem

How might we improve the quality of briefing notes by providing the access to consistent, and timely data and information to analysts?

II. Background

The analysts of NRCan (Natural Resources Canada) often have to produce briefing notes within a very short time (normally ½ to 2 days) for the executives to make important decisions. They usually find themselves spend 60 to 70 percent of the time gathering data and / or information rather than performing analysis and writing the briefing notes. This results in that the insights which they provide could be not in-depth, and not inconsistent. The cause of the issue is that there are a vast amount of both internal and scattered external data sources that the analysts have to search through with limited digital tools, and digital skills.

Another problem is that the analysts move around frequently to different positions, and the average turnover time is 18 months according to the solutions leader, analyst, and stakeholders that we have interviewed. For the new analysts, not only there is a learning curve for subject matters businesses, the training NRCan provides is not specialized enough to allow them master the skillsets needed to write high quality briefing notes in a short time.

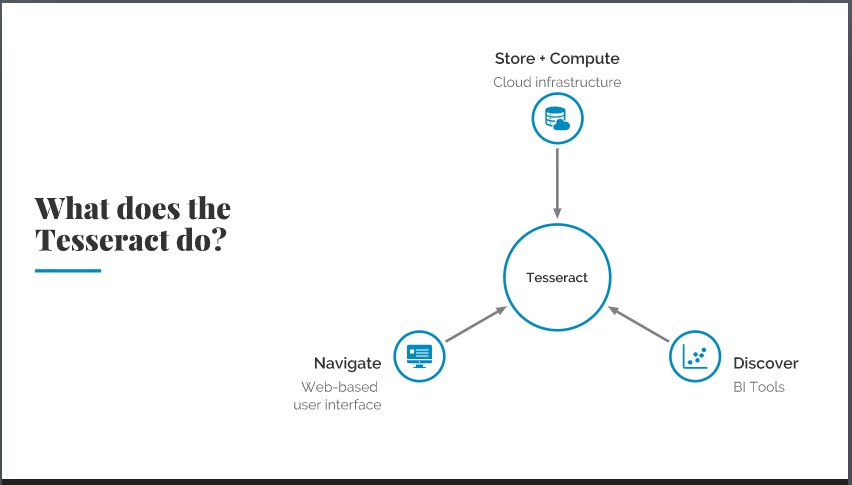
III. Technical Solution

1. **Minimize time gathering information**

The analysts use both microdata and sentiment / text data. They are found from the huge amount of data sources both internally and externally. External data sources are information from the internet, data providers outside of NRCan. They are institutions and organizations such as Bloomberg, Statistics Canada, OECD (Organization for Economic Co-operation), IEA (International Energy Agency), and etc. The Internal data sources is mainly GCDOCs at this time that contains not only the internal data, but also data collected by the analysts from the external data sources. Both internal GCDOCs and external data sources are very hard to use with the search function limitations. To speed up the search and minimize the time to gather information for the analysts is evidently the key to solving the problem.

1. **Tesseract – the Robust Application**

We call this solution Tesseract, because with all the components that are listed below, ideally the application provides the analysts endless power to get data, and extract up-to-date, accurate data and information in 30 to 40 percent of their given time. Ten they can do the analysis and spend time to write briefing notes with consistent and in-depth insights in a timely fashion. It will, without any doubt, help the decision makers to make the best decisions for Canada and Canadians.



1. **Cloud – centralize external data sources**

In order to speed up the data collecting time, we (Team 9 hereafter) designed a system with the following components:

* A centralized data warehouse / storage implemented with Microsoft Cloud Azure to store all external and internal data sources that the analysts are interested, that are subject matter relevant
* Automated and manual processes load the data and information from external and internal data sources into the cloud data storage

1. **Cloud – leverage effective search**

* Azure provides the Search services to index digital data and text information to allow efficient search
* Azure also provides the AI and Machine Learning (ML) services, which can greatly improve the search for relevance, trends, and help with depth of analysis results

1. **Graphic user interface (GUI) – easy access**

To make the solution easy to access and user-friendly, this system will be a web application of a single entry point. It provides web pages to allow the analysts to enter the search criteria to narrow down their search for relevant data and information. The user can access the tool simply by entering a URL on a web browser, and bookmark it for later usage.

The user can select to perform a search of microdata or a search on sentiment information. There are Keywords search function and Advanced search function. When the search results are displayed, there is data filter features to narrow down/organize the search results in difference categories such as News, Government, GCDOCS, Publications, datasets, and Twitter, etc.

The tool is integrated with BI tools to create visualized data reports from the search results. The reports contain both text summaries, numbers, and / or graphics allowing use to perform analytical work, or put them into the briefing notes.

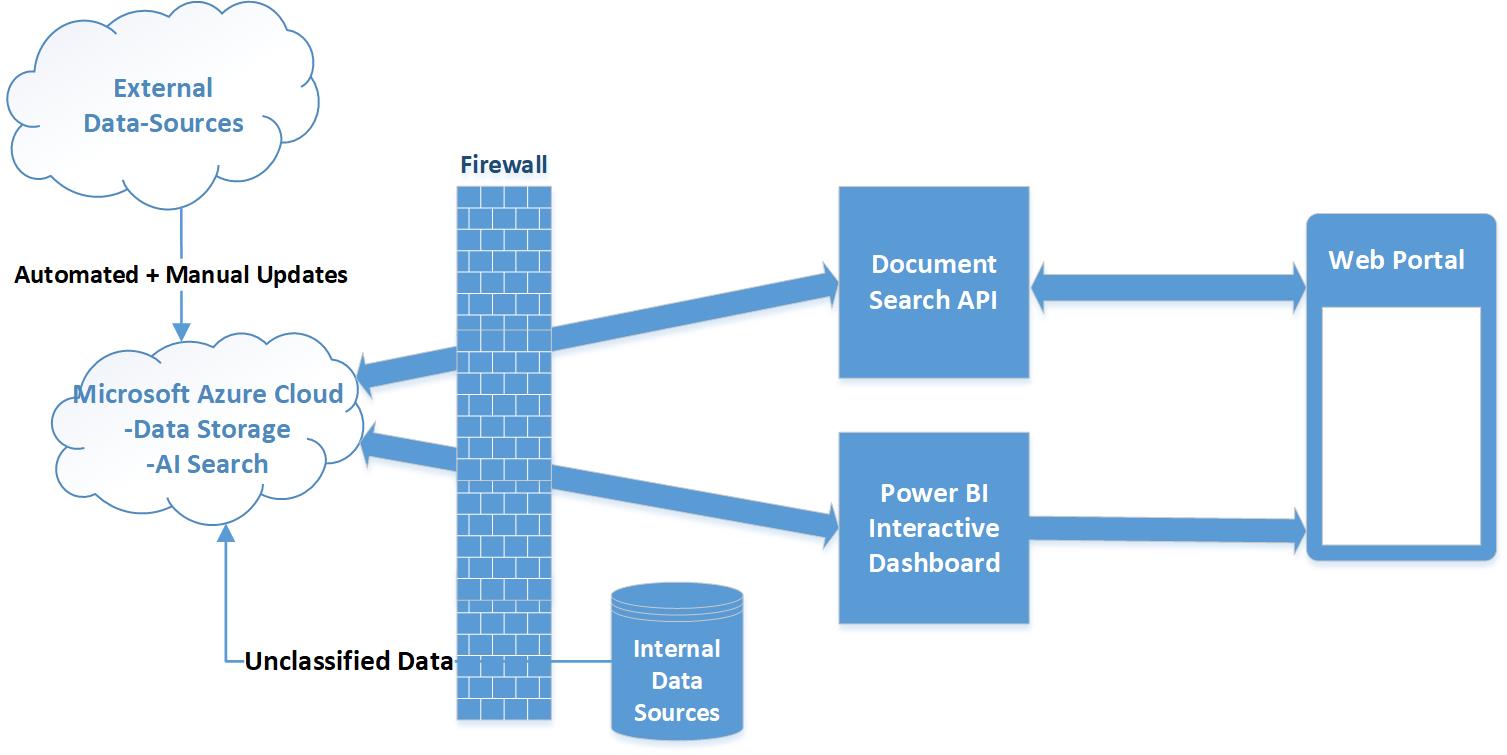
1. **Analysis results - Business Intelligence （BI）Tool**

The above mentioned numbers and graphics are produced from the business intelligence tools (Open source or not) that is integrated into the web application. Once the data are gathered from the search function, the data are feed into the BI tool with pre-loaded custom scripts to generate the interactive graphics on the dashboard.

1. **Architecture Diagrams and Architecture**

Here are the High-level and detailed Architecture diagrams that shows all the components, technologies used, and relationships among the components.

**7.1 High-level Architecture Diagram**



In the High-level Architecturediagram, here are the major components (left to right, top to bottom):

* The external data sources
* Automated and manual processes to centralize data
* Azure data storage (Cloud, index, score, AI search, AI/ML)
* The firewall
* The internal data sources (NRCan GCDocs)
* Document search API
* Power BI interactive Dashboard for digital data
* Web portal with web pages that allow the user to do Keywords, and Advanced search
* Web portal also allows the user to apply filters for search results

**7.2 Detail Architecture Diagram**

**A screenshot of a social media post

Description automatically generated**

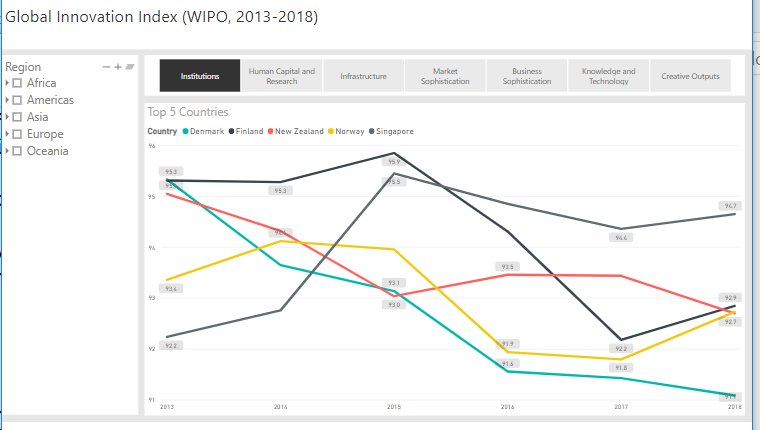
In this Detail Architecture Diagram, more details are included to explain the not only more relationships among the components, it also listed the technologies and services in each further broken-down modules.

* The application / system starts by feeding various data sources (Internal data sources, external data sources that are documents and databases) into a centralized Cloud storage using Automated and / or Manual Processes into AI component.
* An AI component is implemented with Python scripts to generate quotes, keywords, and summaries from documents and articles and load into the Azure Cloud Storage
* Power BI (can be Open Source BI tools) that creates interactive visual presentation of digital data in the Clod Storage and present on the Web application’s front-end
* The Azure Cloud Storage is the centralized data store that holds both the un-structured semantical as data Blobs and the structure data Tables from all data sources
* On the data processing side, the Tesseract’s Web Services are implemented using Cloud / Azure services / components to provide functionalities for the web application’s front-end, they are:
  + Cloud / Azure Cognitive AI, including Multilingual, Translation, Text Analysis, and Sentiment Search
  + Cloud / Azure Search, including Azure Search, Indexing
  + External API’s Data sources, including Being Search, Google Patents, Twitters Feed, and LinkedIn
* The Web Application front-end, a single entry point for the user with a URL

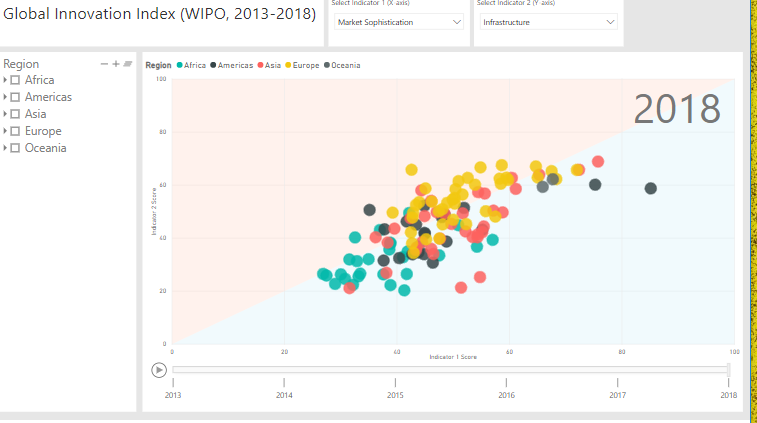
1. **Power BI and dashboard**

Power BI’s dynamic, interactive dashboard that shows the reports of digital data.

**8.1 Fig 1 for dashboard graphics – line graphes**



* 1. **Fig 2 for dashboard graphics – scatter plots**

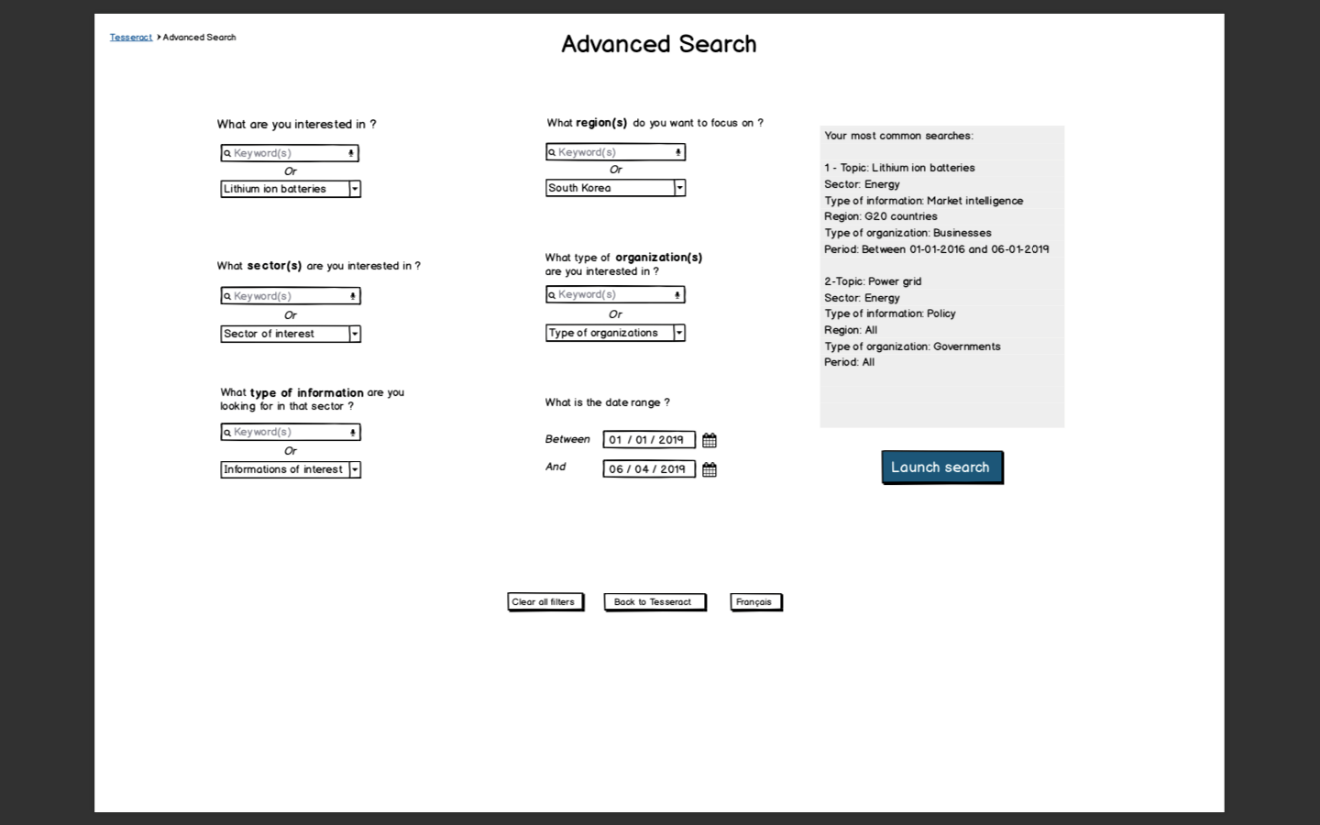


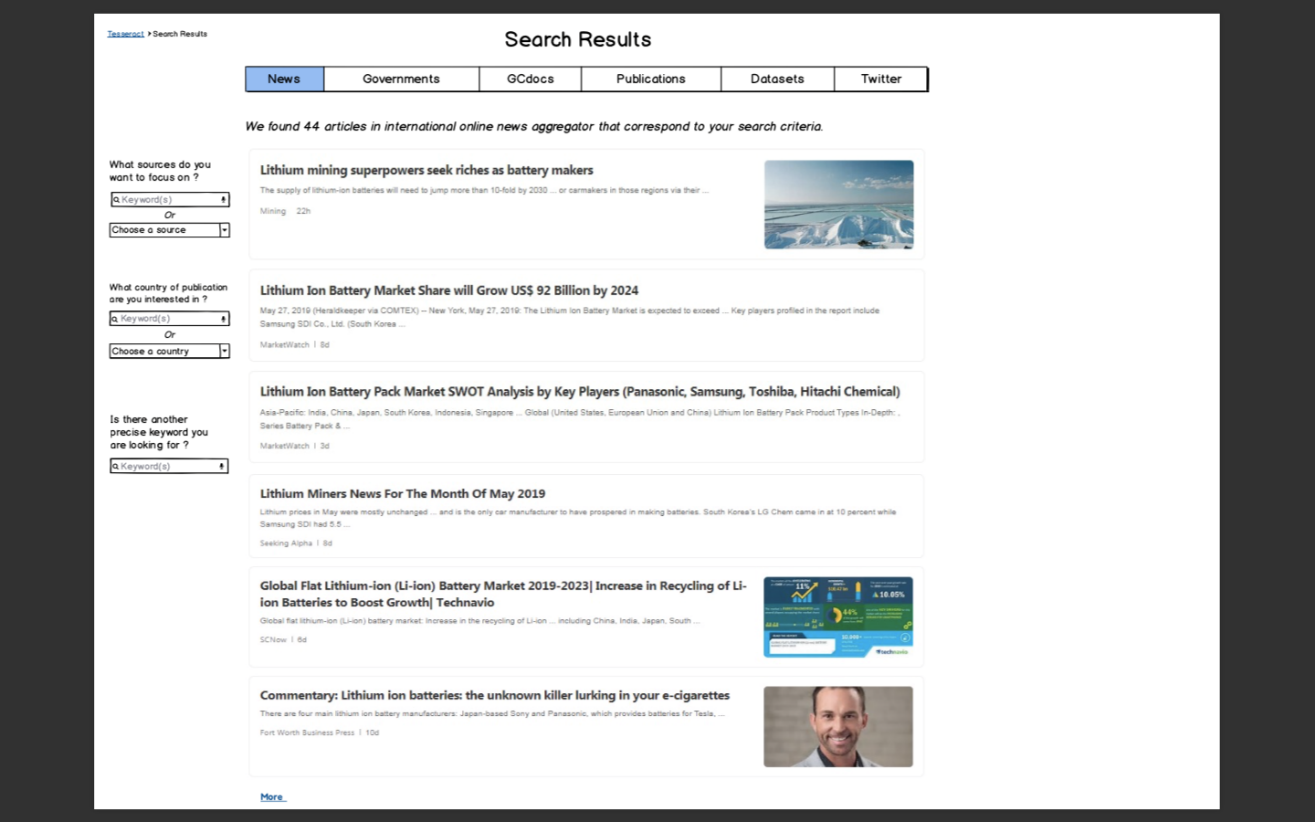
1. **Web portal**

This is the tools presentation component including web pages, and dashboard mentioned above. Here are the web pages’ screenshots:



Secreenshot1: This page allow the user to do keywords search of sentiment data or see General Trends of structured data from BI dashboard; or doo Advanced search.

Screenshot 2: Advanced search, allow the user to enter criteria for advanced search.



A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a computer

Description automatically generated

Screenshot3 to 8: Filter search results by news, or Government, GCDocs, Publications, Datasets, or Twitter.

IV. Recommendations

1. **Internal data sources**

Improve the GCDOCs of NRCan search functions to allow strong search features, performance and relevancy on up-to-date data and semantical information.

Provide with analysts with analytical tools that needed to get better insights from the search results.

1. **External data sources**

To continue implementing the full designed features as described in the section of Technical solution.

1. **Subject matter specialist network**

Include in the system, or in GCDOCs of NRCan, an easy accessible list of contact information of subject matter specialists. This is important because it has been the most used shortcuts for getting valuable, relevant, most up-to-date and consistent insights on subject matters by the analysts.

1. **Expand scope of the tool**

Right now the end results is to minimize the data gathering time and provide easy accessible web application and BI tool. In the future the end results of the tool can be briefing notes itself with dynamic visual representations on top of the traditional briefing notes. This will extend/strengthen the executives convincing power for those who have the capability /training to illustrate the data and graphics.

For future development on this project, here is the code repository:

<https://github.com/marwanfarah/accelerate-evidence>

1. **Human resources**

**Hiring** - Consider / provide funding to hire more CS category employees to satisfy the technical needs for automate the requirements of the analysts.

**5.1 Analytic skillsets** - Facilitate regular training for the new employees in more specific / in-depth of subject matter skills. In the meantime setup long-term collaboration with other government departments who have the already regular training by their specialists and expertise.

**5.2 Digital skillsets** - Include the more digital skillsets in analysts’ mandatory training. In this fast evolving digital world, the abilities to use data analytical, business intelligence tools, to write basic scripts to help analyze data have become part of the analysts daily life.

Only analysts possess both analytical and digital skillsets with good subject matter knowledge can produce the effective briefing notes efficiently.

V. Prototype

The purpose of this prototype is to implement a working application for prove of concept, to convince the executives and analysts to accept it.

Due to the constraints on time, resources, and complexity of the system, we have narrowed down the scope to lithium-ion batteries datasets, keywords search and limited search results filter functions. Prototype features details are described as the followings:

1. **Web user interface (GUI)**

* Easily accessible by entering a URL on web browser
* Web GUI that is bilingual and conforms the Web Content Accessibility Guidelines
* Keywords search functions and present the sentiment search results

1. **Data and centralized storage**

* Data is limited to external, public data - digital data and text data on lithium-ion batteries
* Microsoft Azure Cloud data storage for text / sentiment data

1. **Sentiment Search features**

* Azure search services for text data - index, score for relevance and translation features of limited languages
* Python scripts using AI / ML features to generate summaries from text contents
* Expert Data sources such as Patents are included in the initial datasets of the document search

1. **Structured data reporting**

* Power BI tool for digital data - customized R scripts to produce a dashboard that presents the dynamic and interactive graphs of reports

1. **Tools and Technologies used**

* Microsoft Azure, Power BI
* GitHub, Visual Studio, and etc.
* Python, .net, R, HTML programming / scripting languages
* AI / ML API, Open Source APIs such as Google Patent Search API

VI. Future implementation

To promote Open sources within the GC, the code for the prototype is stored in GitHub. It is readily to be shared by NRCan for further implementation. The prototype also utilize the most trendy, optimal modern technologies such as Python, Cloud (Azure), Web / Internet technologies, powerful business intelligence as data analytical tools, and dashboard for data and trend / prediction with visualization. It allows the future full implementation to be on the right direction for security (Cloud security services), performance, for the conforming the industry stands of WCAG (Web Content Accessibility Guideline) and user experiences.

1. **Features include**

* All external and internal, public and protected, digital and text data of stakeholders’ interest
* Build in security for protected data and document contents
* Cloud Azure efficient search functions with AI / ML features enhancing/speed up text data search
* Design efficient central data warehouse for the microdata
* Integration of Web application with business intelligence tools and dashboard to allow the dashboard dynamically interact with datasets and semantical information of the same topic (search results from the same keywords)
* Web pages fully bilingual, accessibility compliant, good user-experience GUI
* Build the complete clean technology taxonomy, and other taxonomies required into this tool

VII. Appendix

**Stakeholders**

Annie Prigge

Deputy Director – Asian Pacific, Policy and Planning,

Innovation and Energy Technology NRCan

**Solution leader**

Jack Jensen

Combustion specialist, electronics engineer and researcher

Science & Technology Advisor – Policy Analysis and Coordination, Policy and Planning, Innovation and Energy Technology NRCan

**Interviewees**

Jack Jensen, Annie Prigge, Qi Wang

**Team 9 – Avengers**

(In order of the last name)

Julien Aubin-Beaulieu

Marwan Farah

George Flerovsky

Eugenio Lo

Timothy McLarty

Bo Peng

Gillian Whichelo

VIII. Acronyms

AI - Artificial Intelligence

BI – Business Intelligence

GC – Government of Canada

GUI – Graphic User Interface

ML – Machine Learning

NRCan – Natural Resources Canada

URL – Uniform Resource Locator

WCAG - Web Content Accessibility Guidelines

References

<https://www.w3.org/WAI/standards-guidelines/wcag/>

<https://opensource.com/resources/what-open-source>

<https://docs.google.com/document/d/1k02HTkK2rDYclTNtT7DSaHA5p6sLidppVoGKk_TpQg8/edit?usp=sharing>

<http://publications.gc.ca/site/eng/home.html>

<https://batteryuniversity.com/learn/article/bu_1101_glossary>

<https://open.canada.ca/data/en/dataset/c1b0f627-8c29-427c-ab73-33968ad9176e>

<https://www.wipo.int/publications/en/details.jsp?id=4330>

<https://batteryuniversity.com/learn/article/bu_1101_glossary>

<https://energy.economictimes.indiatimes.com/tag/bloomberg+new+energy+finance>