THOMSON REUTERS

mla donor COMPANY

final report

Value Adding

Project code: **P.PSH.0693**

Prepared by: **Russel Rankin**
Food Innovation Partners Pty Ltd
Russel@food-innovation.com.au

Andrew O'Brien
Thomson Reuters
andrew.obrien@thomsonreuters.com

Date: **April 2015**

Identifying agrifood opportunities for designing food products for Healthy Ageing – a Cortellis data fusion pilot for mining big data.

Meat & Livestock Australia acknowledges the matching funds provided by the Australian Government to support the research and development detailed in this publication.

P.PSH.0693 – Identifying agrifood opportunities for designing food products for Healthy Ageing – a Cortellis data fusion pilot for mining big data.

April, 2015

Project Leader: Russel Rankin
Food Innovation Partners Pty Ltd
0411 178 227
Russel@food-innovation.com.au

Purpose: Undertake a pilot project using Thompson Reuters, Cortellis Data Fusion to analysis big data for the purpose of identifying agrifood opportunities for designing food products for healthy ageing. A pilot project allowed the meat industry to evaluate the effectiveness of data analytic tools to identify trends and opportunities.

Acknowledgment of funding sources:
Food Innovation Partners, Thomson Reuters and The funding establishment acknowledge the financial support for this project from Meat & Livestock Australia (MLA), the Meat Donor Company (MDC) and Dairy Australia Limited.

Abstract

Food Innovation Partners (FIP) and MLA have identified a need in the food industry for capabilities provided by Thomson Reuters in the analysis and reporting on 'Big Data' for inclusion in the MLA's Global Insights Program.

The funding establishment have identified a market opportunity in the market segment of foods for healthy ageing. In order for The establishment to evaluate the market opportunity they need to gain deep insight into the market segment in order to make better informed strategic decisions.

As a proof of concept, Thomson Reuters were engaged in a services based project to prove the capabilities of their 'Cortellis Data Fusion' platform as the data management layer technology that can;

i). Identify relevant trends in the area of 'Foods for healthy ageing' that The establishment have the potential and capability to take to market, and

ii). Enable MLA to assess the value that Big Data analysis can provide the Australian food industry in the current context of masses of literature, patents and social media commentary.

The analysis included a primary data layer; the MLA and DA data store related to 'Foods for healthy ageing' currently hosted in a Dropbox folder. It also included a second layer of data including the Web of Science and a third layer 'social media'.

This project provided an opportunity for FIP, MLA and The establishment to evaluate an approach to mining big data and to interrogate previous research findings, trends and insights to inform strategic directions by learning from data.

Executive summary

Ageing is determined by complex interactions between biological, environmental, socioeconomic, and cultural factors, some of which are beyond the control of individuals. A number of external factors contribute to the ageing process, such as poor nutrition, physical inactivity, smoking, and psychosocial characteristics (such as stress). These factors are associated with the development of chronic diseases that are, in themselves, associated with physical and mental frailty and could be tackled at an individual level throughout life.

- Ageing affects people in different ways, with a wide variation in age related physical and mental functioning
- Healthier ageing is achievable through modifying some lifestyle factors—such as stopping smoking, being more physically active, and eating a balanced diet
- For healthier ageing, eat mainly nutrient dense foods that are rich in vitamins and minerals and low in fats and sugar
- Balancing energy intake and expenditure is important for maintaining healthy weight

Diet is arguably the single most important behavioural factor that can be improved which will have a significant impact on health. The quality and quantity of foods and drinks consumed has a significant impact on the health and wellbeing of individuals, society and the environment; better nutrition has huge potential to improve individual and public health and decrease healthcare costs.

Identifying emerging trends and opportunities in the market, amongst the huge quantities of market, technical, consumer and social data is becoming more difficult as the volume of information increases. Food Innovation Partners (FIP), MLA and DA have identified a need for capabilities provided by Thomson Reuters in the collation, curation, analysis and reporting on 'Big Data' for inclusion in the MLA's Global Insights Program. Current methods for analysing data to gain insights relies on individuals reading the data then determining insights that can be precursors to emerging trends. The challenge for these manual methods is that different insights can be gained, depending on who reads, analyses and interprets the data. The food industry along with many other sectors is seeking new technology based methods for analysing large quantities of information, repeatably.

The funding establishment has identifed an opportunity to develop a range of meal solution products targeted at consumers 60+ years of age. The funding establishment is working with MLA and DA and are seeking a partner to support the compilation and analysis of Big Data content requirements for the identification of opportunities for innovation in the 'Foods for health ageing' market segment.

As a proof of concept, Thomson Reuters was engaged to undertake a pilot project to prove the capabilities of Cortellis Data Fusion as the data management layer technology to;

iii). Identify relevant trends on the area of 'Foods for healthy ageing' that The establishment have the potential and capability to take to market, and

iv). Enable the MLA and DA to assess the value that Big Data analysis can provide the Australian food industry in the current context of masses of literature, patents and social media commentary.

The analysis included a primary data layer consisting of market, consumer and industry reports, aged population health studies and analysis related to the 'Healthy ageing' market segment. It also included a second layer of data including the Web of Science and FSTA and a third layer, 'social media'.

This initial pilot project has demonstrated that there is a need for Big Data analytics tools in the food industry going forward in order for commercial food companies to be able to identify commercial opportunities based upon global market and consumer insights.

The pilot project identified a number of critical issues that need to be addressed in order for data analytic tools to be effective in the food industry and some of these are listed below;

i. Access to primary or raw data in a format that can be accessed by Cortellis tool.

ii. Optimal data format is RDF (Resource Description Framework). Where this is not possible then a flattened CSV or Excel file is acceptable. Data should be reviewed by TR's CDF team prior to project work commencing.

iii. Data needs to be formatted into a structured contextual foundation in which to build the core semantic framework to allow the unstructured documents to stitch to relevant entities for analysis. By age group, by food products, etc (FSTA Thesaurus).

iv. Data formatting and review must be factored in to overall project timelines

v. Agreed question structures, ontologies, search terms and primary keys are critical to successful data modelling. Thomson Reuters has Food & Beverage expertise through their Scientific & Scholarly Research team that can be used in an advisory capacity for development of ontologies and other areas requiring domain expertise.

vi. Data should entail some curation for best integration to answer questions

vii. Unstructured data provided limited in functionality, i.e. tables in pdfs unreadable without the raw data in current instance but able to build into full development plan.

The initial Pilot Project activities conducted with MLA, DA and The establishment developed the following recommendations;

1. Thomson Reuters undertake a 2nd Pilot Project to provide The funding establishment with the required answers to their queries related to 'Foods for healthy ageing' and to demonstrate the effectiveness of TR's Cortellis Data analytics platform on Bid Data in the food industry. The initial pilot allowed Thomson Reuters to gain huge knowledge related to the quality and format of information available in the food and health sector and how to most effectively and efficiently analyze the data.

   Unfortunately the pilot study did not provide The establishment with the market and consumer insights that was anticipated. Thomson Reuters believe that their Cortellis tool can deliver the big data analytics capability that MLA, DA and The establishment desire. Thomson Reuters have offered to continue work on the Pilot project with The establishment in the 'Foods for healthy ageing' sector (at no additional charge) in order to demonstrate to all stakeholders that their Cortellis analytics tool can deliver the anticipated insights.

2. Conduct a Phase II project with MLA, DA and The establishment with a Define & Design component in order to review and hone the needs of MLA's Global Insights program (including structure). This phase will need to be budgeted and include appropriate Go/No-go milestones. Phase II would involve building the visualisation tools and would involve the following activities;

   - Review of data sources and structure for improved integration.
   - Propose, review and adapt as necessary required ontologies and search criteria.
   - Review curation requirements and secure resource to ensure data integrity.
   - Review and agree data modeling approach to answer specified questions.
   - Ensure data visualisation and end user "usability" aligns to MLA, DA and The establishment's needs.
   - Define Workflows to answer key questions.
   - Phased roll out to MLA Global Insights program.

Contents

Background

Ageing is determined by complex interactions between biological, environmental, socioeconomic, and cultural factors, some of which are beyond the control of individuals. Factors that contribute to the ageing process, such as poor nutrition, physical inactivity, smoking, and psychosocial characteristics (such as stress), may be modifiable. These factors are associated with the development of chronic diseases that are, in themselves, associated with physical and mental frailty and could be tackled at an individual level throughout life.

- Ageing affects people in different ways, with a wide variation in age related physical and mental functioning
- Healthier ageing is achievable through modifying some lifestyle factors—such as stopping smoking, being more physically active, and eating a balanced diet
- For healthier ageing, eat mainly nutrient dense foods that are rich in vitamins and minerals and low in fats and sugar
- Balancing energy intake and expenditure is important for maintaining healthy weight

Diet is arguably the single most important behavioural risk factor that can be improved which will have a significant impact on health. As the quality and quantity of foods and drinks consumed has a significant impact on the health and wellbeing of individuals, society and the environment, better nutrition has huge potential to improve individual and public health and decrease healthcare costs.

The funding establishment has identifed an opportunity to develop a range of meal solution products targeted at consumers 60+ years of age. These products would either be provided to institional care and meals on whells' or through regular retail channels. A proposed product range would potentially be suitable for both domestic and export markets. Succesfull development of meal solutions for 60+ market segment has the potential to increase signfiicant demand for both red meat and dairy protein through market uptake in Australia and South east Asia.

The intial activity for The establishment is to identify the key trends and opportunities within the segment labeled 'Foods for Helathy Ageing'. This project investigated the effectiveness of Big Data analytic tools by conducting a pilot project using Thompson Reuters Cortellis Data Fusion.

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set.

Every 10 minutes more data is generated than was generated over the last 70 years. In 1992, 100 GB of data was created every day. Today that is 29,000 GB per second and in 2018, 50,000 GB per second. In the food industry this data includes research reports, analysis, trademarks and IP, social media and many more data sets.

Current methods for analysing data to gain insights is for individuals to read the data and determine insights that can be precursors to emerging trends. The challenge for these manual methods is that different insights can be gained, depending on who reads, analyses and interprets the data. The food industry along with many other sectors is seeking new technology based methods for handling, analysing, capturing, curating, searching, sharing and, storing. MLA has identified the urgent need to investigate these types of data analytic

tools to help improve the effectiveness, efficiency and speed of gaining market and consumer insights.

Thomson Reuters Cortellis™ Data Fusion (formerly Entagen's Extera) enables firms to gain new scientific and strategic insights by connecting a firm's proprietary content with the world's data, utilizing Big Data and Linked Data technologies. A powerful proprietary data core technology is used to aggregate entities and create associations between them, enabling organizations to securely integrate subject matter content from a variety of sources. Cortellis Data Fusion can identify 'hidden' relationships between data entities that may be prcursors to trends and emerging opportunities. Thomson Reuters Cortellis platform is currently used by top pharma and research institutions for applications such as target finding, drug repurposing, and precision medicine. This pilot project is the first attempt to apply their data anlaytics technologies to information and data in the food industry.

TR's Cortellis Group has develop web based visualizations to allow users to create, search & share structures; knowledge "maps" of associations to help uncover unexpected associations, generate hypotheses and predict emerging trends.

Project objectives

The following key outcomes were planned to be delivered:

1. A report on data integration within Thompson Reuters' Cortellis Data Fusion platform using "Healthy Ageing" data content provided by MLA, The establishment and Food Innovation Partners to support The establishment's review and development of the market opportunity of 'foods for healthy ageing', and
2. Demonstration to stakeholders on how Thompson Reuter's professional services data modelled the content and example dashboard interface capabilities utilizing Cortellis Labs development resources in identifying insights into Healthy Ageing and Food Design.

Methodology

This project involved undertaking the following methodology;

Collation of existing information and data.

FIP worked with MLA, DA, HIA and The establishment to collate existing information relating to the "Healthy Aging" sector. This data was a primary data layer consisting of market, consumer and industry reports, aged population health studies and analysis related to the sector.

During this stage the project design and expected outcomes were defined with all stakeholders.

Understand the questions being asked of the data

Reviewed how the end user (MLA and The establishment) wanted to navigate/interact with the data. Developed the entities that would support the relationships needed to help answer the questions

Checking information and data format and structure

TR reviewed the data collected in the project Dropbox for structure and format and began to compile relevant information that they had access to such as FSTA (Food Science and Technology Abstracts), World Patent Index etc. The project team also reviewed the data structure(s) of interest and began to organize/edit for best integration.

TR and The establishment define questions for analytics tools

Thomson Reuters met with The funding establishment and Food Innovation Partners to develop a series of queries to apply to the data set 'Healthy Ageing' using the Cortellis analytics tool. Below are some of the queries that would provide The establishment with some valuable insights.

- Top 10 health concerns (+/- changes), list rank and data table?
- Key 'health themes' from the data?
- Food preferences, attitudes etc by occasion by age - Of the different age splits, how do attitudes, health issues, behaviour, household structure, ethnicity or food attitudes change?
- Is there synergy with government legislation, vs the key themes and consumer trends? Are there areas the government is not including that The establishment could lead the agenda?
- What are the demographic, life-stage trends and differences between them?
- Are there any lifestyle, age or attitude differences dependent on the disease state of an individual?
- What are the most prevalent frustrations or 'fears' in food and lifestyle dependent on the disease state an individual has?
- Differences between sensory - taste, packaging, texture; by age or disease state?
- What proportion of this market is taking dietary supplements? What are they taking?
- What is the ideal level of protein intake as one ages?

TR compiled data for connecting together with in the Cortellis data platform.

TR developed a data modeling plan based on 'stitching' rules to connect the data sources to the Cortellis platform in order to run analytics. TR then ingest data and model based on the developed 'stitching' rules.

Search strategy

The search strategy involved a search of key scientific databases related to nutrition, functional foods and ageing. Six key topic searches were performed separately using predetermined food ingredient search terms. All studies were limited to persons over 40 years of age. For the purpose of this ingredient-based review and due to the size of the research area, only high quality studies using controlled clinical trials, randomised controlled trials and meta-analyses were included. These would traditionally be assigned the greatest weighting in evidence tables. Identifying studies involved searching electronic databases (Medline, CINAHL and Science Direct).

Data analytics reporting.

TR conducted a Workshop and a series of meetings with The establishment and FIP to report progress during the Pilot Project and to present the final outcomes of the pilot work. A copy of the presentation is included in Section 7.

Results and discussion

The table below provides details of the outcomes of the analytic queries; What data is available? What are the data gaps? What could be done to answer The establishment's questions?

Table 1: 'Foods for healthy ageing' data sets and integration requirements.

KEY QUESTIONS								
	Question	**Answer status**	**Required data integration**	**CDF Value**	**Data summary**	**Data Gap**	**Data mapping**	**Manual data review requirement**
1	Top 10 health concerns (+/- changes), list rank, data table	Answered	no	Very easy	Health data			
2	Key 'health themes' from the data	Part answered	maybe	OK example	Map health data to health themes data	Needed to define themes		
3	Food preferences, attitudes etc by occasion by age - Of the different age splits, how do attitudes, health issues, behaviour, household structure, ethnicity or food attitudes change?	Not answered	yes	OK example	Map food preferences data to health data		Needed to think more about id's, as couldn't map health data with food preferences data. We could think implement a series of workarounds, either when loading data or at the visualisation level	
4	Is there synergy with government legislation, vs the key themes, consumer trends big data uncovers - are there areas the government are not including that we could lead the agenda?	Not answered	yes	Unknown	Wasn't sure of the best data.	Needed to define themes - what are these? They would be needed to be mapped to the docs beforehand. Wasn't sure about the legislation docs		potentially

5	What are the demographic, life stage trends and differences between them	Answered	no	Very easy	Popn data can be used. But we need better visuals to show off the trends.			
6	Are there any lifestyle, age or attitude differences dependent on the disease state of an individual	Answered	yes	Good example	Popn trends data and bioimarker data			
7	What are the most prevalent frustrations or 'fears' in food and lifestyle dependent on the disease state an individual has?	Not answered	unknown - docs	Bad example	Wasn't sure about the data. Believe the relevant information was buried in text docs. Needed an analyst to identify the relevant docs then review.	Did we have relevant papers?		yes
8	Differences between sensory - taste, packaging, texture - by age or disease state	Not answered	unknown - docs	Bad example	Wasn't sure about the data. Believe the relevant information was buried in text docs. Needed an analyst to identify the relevant docs then review.	Did we have relevant papers?		yes
9	What proportion of this market is taking dietary supplements? What are they taking?	Answered	no	Very easy	Captured in the supplements spreadsheet			
10	What is the ideal level of protein intake as one ages?	Not answered	unknown - docs	Bad example	Wasn't sure about the data. Believe the relevant information was buried in text docs. Needed an analyst to identify the relevant docs then review.	Did we have relevant papers?		yes

Table 2: Lessons learnt versus desired data queries.

Priority	Name	Description	Lessons Learned/Deliverable
1	Unmet needs	Identification of a number of clear unmet consumer needs in foods for +60 years old. (Example Complete meal, fresh food, fortified meals etc)	Raw data embedded in PDFs is not available to be surfaced and incorporated into analyses. Therefore embedded data would preferably be provided in an alternative format for consumption by CDF instance The structure and specificity of the question is critical to the success of cross document linking and consolidation of data. TR shall need to work with The establishment to establish appropriate search criteria/ phrases and Primary Keys to be used as document "connectors" and/ or identifiers
2	Needs ranking/weighting	Weighting or ranking of unmet consumer needs in foods for +60 years. (Ranked in order of importance. Maybe weighted)	Existing CDF interface is designed for data linking and is not designed for visualisation or analytics. TR is able to implement an API (Analytics Programme Interface) to export the relational data identified in CDF into a third party software such as Tibco Spotfire. Here TR is able to develop custom analytics dashboards to aide data visualisation and ranking in-line with The establishment specifications In the current instance it is possible to sort data on numeric values - a more basic approach to the above.
3	Barriers/impediments	Identification of barriers and impediments to consumption. (taste, price, convenience, ease of opening etc)	The structure and specificity of the question is critical to the success of cross document linking and consolidation of data. TR shall need to work with The establishment to establish appropriate search criteria/ phrases and Primary Keys to be used as document "connectors" and/or identifiers of relevant information.
4	Functional foods/ Health concerns	What are greatest health concerns related to food and nutrition for the +60 YO group?	The structure and specificity of the question is critical to the success of cross document linking and consolidation of data. TR shall need to work with The establishment to establish appropriate search criteria/ phrases and Primary Keys to be used as document "connectors" and/or identifiers of relevant information

			TR is able to implement an API to export the relational data identified in CDF into a third party software such as Tibco Spotfire. Here TR is able to develop custom analytics dashboards to aide data visualisation and ranking.
5	Product ideas/ Formats	New product ideas. New product delivery platforms/presentation formats/Form and function.	The structure and specificity of the question is critical to the success of addressing a specific question. In this example, CDF is not “intelligent” enough to understand the objective, link and return appropriate data to identify and validate areas of opportunity. TR recommends that the appropriate approach to this type of analysis is to use CDF capabilities to complete a gap analysis, based on agreed assumptions. The output of this can then undergo human validation, either via The establishment, TR internal resources or industry experts through approaches such as primary market research. Approach carried out by TR for pharma industry. Potential to develop heat maps (examples to follow) to easily identify potential market gaps.
6	Size and value	Indication of size and value of meal solution market total and by meat, dairy component.	The output of this can then undergo human validation, either via The establishment, TR internal resources or industry experts through approaches such as primary market research.

Conclusions and Recommendations

This initial pilot project has demonstrated that there is an opportunity for the application of Big Data analytics tools in the food industry. There are critical issues and challenges that must be addressed in order for data analytic tools to be effective in the food industry and some of these are listed below;

i. Access to primary or raw data in a format that can be accessed by Cortellis tool.

ii. Optimal data format is RDF (Resource Description Framework). Where this is not possible then a flattened CSV or Excel file is acceptable. Data should be reviewed by TR's CDF team prior to project work commencing.

iii. Data needs to be formatted into a structured contextual foundation in which to build the core semantic framework to allow the unstructured documents to stitch to relevant entities for analysis. By age group, by food products, etc (FSTA Thesaurus).

iv. Data formatting and review must be factored in to overall project timelines

v. Agreed question structures, ontologies, search terms and primary keys are critical to successful data modelling. Thomson Reuters has Food & Beverage expertise through their Scientific & Scholarly Research team that can be used in an advisory capacity for development of ontologies and other areas requiring domain expertise.

vi. Data should entail some curation for best integration to answer questions

vii. Unstructured data provided limited in functionality, i.e. tables in pdfs unreadable without the raw data in current instance but able to build into full development plan.

This report makes the following recommendations;

1. Thomson Reuters undertake a 2nd Pilot Project to provide The funding establishment with the required answers to their queries related to 'Foods for healthy ageing' and to demonstrate the effectiveness of TR's Cortellis Data analytics platform on Bid Data in the food industry. The initial pilot allowed Thomson Reuters to gain huge knowledge related to the quality and format of information available in the food and health sector and how to most effectively and efficiently analyze the data.

   Unfortunately the pilot study did not provide The establishment with the market and consumer insights that was anticipated. Thomson Reuters believe that their Cortellis tool can deliver the big data analytics capability that MLA, DA and The establishment desire. Thomson Reuters have offered to continue work on the Pilot project with The establishment in the 'Foods for healthy ageing' sector (at no further charge) in order to demonstrate to all stakeholders that their Cortellis analytics tool has application in the food sector.

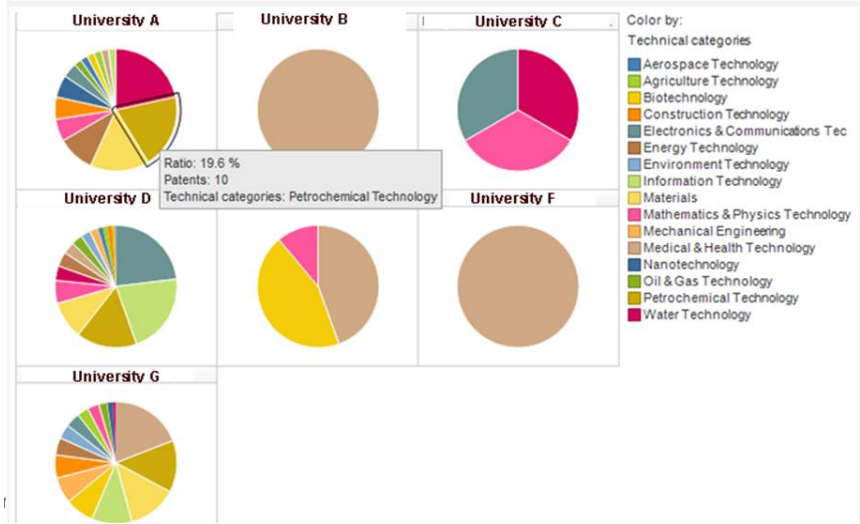
2. Conduct a Phase II project with MLA, DA and The establishment with a Define & Design component in order to review and hone individual questions raised by The establishment (including structure). Phase II would involve building the visualisation tools and would involve the following activities;

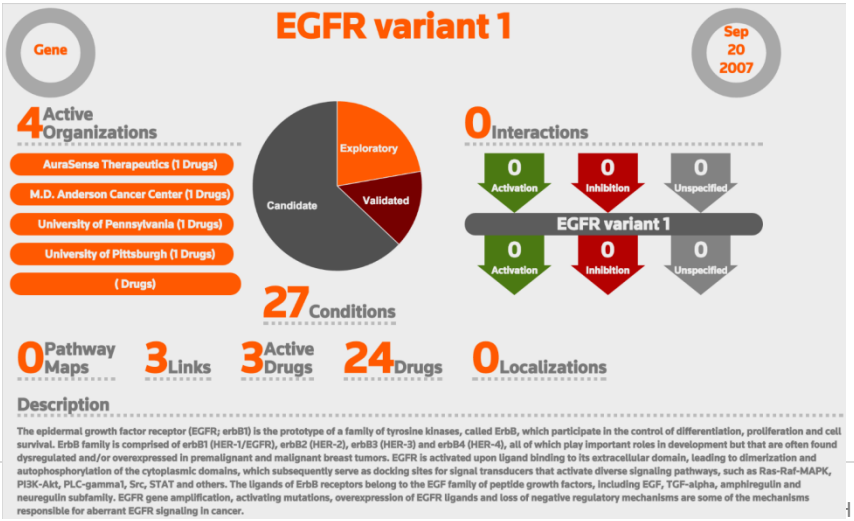
   - Review of data sources and structure for improved integration.
   - Propose, review and adapt as necessary required ontologies and search criteria.

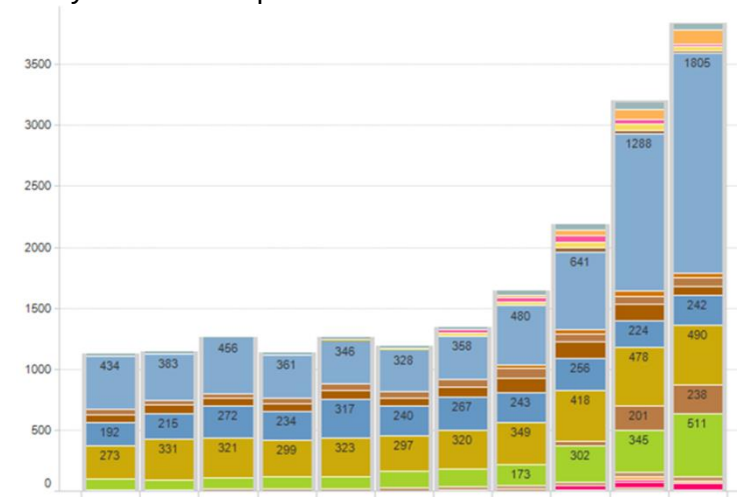
- Review curation requirements and secure resource to ensure data integrity.
- Review and agree data modeling approach to answer specified questions.
- Ensure data visualisation and end user “usability” aligns to MLA, DA and The establishment’s needs.
- Define Workflows to answer key questions.
- Phased roll out to MLA Global Insights program.

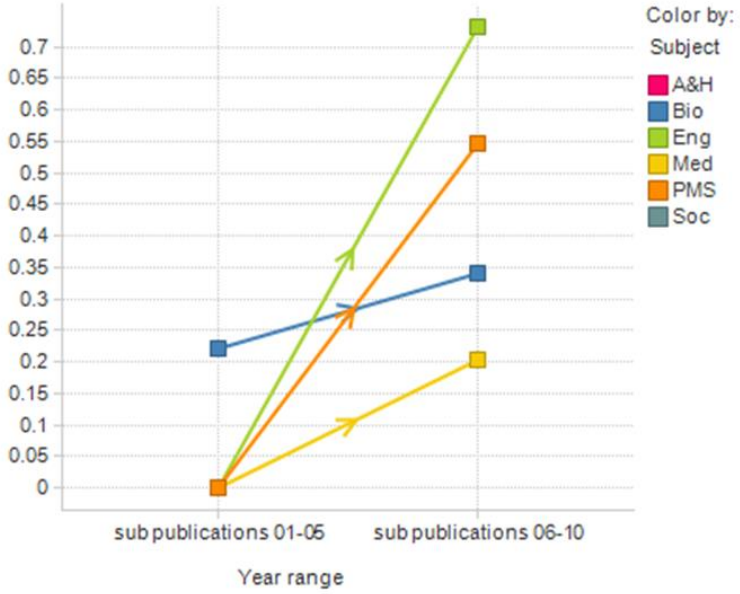
Potential Big Data Visualisation tools

Below is a representation of the visualisation tools that TR can develop to overlay the Cortellis platform.









1. Funding data. Both direct and strategic funding are displayed.
2. Total patents & publications.
3. Staff Student numbers. Staff numbers are broken down by faculty type.
4. Publications by year since 2000.
5. Patents by year since 2000.
6. Publications split by field.
7. Patents split by classification.