Design Document

For

Bayer META (Message Exchange Text Analytics)

Prepared by:

Maggie Lu

Siddharth Menon

Olivia Zheng

Yiling Zhong

Fall 2018 Master of Information Systems Management (MISM) Capstone Project, Heinz College, Carnegie Mellon University

Table of Contents

1. Vision

1.1 Overview

Bayer, like many other global corporations, faces the issue of information segregation in its day-to-day operations. As a result, there may be duplication of effort where multiple teams work on similar/same projects in parallel without sharing domain knowledge or collaborating with each other.

In order to synergize information from different groups and promote knowledge sharing as well as effective collaboration, our team is tasked to create a stand-alone web application that is able to examine the content of emails as well as other textual content in order to connect relevant teams/people together. META, which stands for “Message Exchange Text Analytics”, should be able to utilize Natural Language Processing (NLP) algorithms that identify and cluster keywords. It will generate lists of keywords and send emails to relevant parties.

META should also be able to find and present internet resources which may contain information about the conversation subject. Additionally, the algorithm should be able to take the email recipients’ feedback as input to iteratively update its model. Another important feature would be to allow administrators to edit the lists of keywords due to Bayer’s security and privacy practices.

1.2 Positioning

1.2.1 Problem Statement

The problem of	Siloed communication among different teams and departments
Affects	All who work on innovative projects

The impact of which is	Lack of collaboration resulting in unachieved potential in productivity
A successful solution would be	A tool that can synergize information and model topics from email correspondence, connect relevant parties with useful information and provide the opportunities to collaborate.

1.2.2 Product Position Statement

For	Bayer scientists, technicians, as well as business professionals
Who	Work on multiple projects across multiple functional departments
Our solution	Would connect previously isolated parties with keywords from their email correspondence,intranet knowledge base, and academic writings as well as provide relevant internet resources
That	Promote effective and efficient collaboration

1.3 Stakeholder Descriptions

1.3.1 Stakeholder Summary

Stakeholders	Description	Responsibilities
Business Analysts	Direct Users and Administrators	Provide user requirements Determine metrics of success

The broader Bayer community	Indirect Users	Provide email correspondence and knowledge base from which keywords are found and topics are modeled
The Project Team	Architects, Developers, Testers	Meet the requirements from direct and indirect users Ensure that the project is easily maintainable and scalable for Bayer's own network environment

1.3.2 User Environment

Every night, the data in structured text form would be collected and combined into one text file. The data is then cleaned and transformed into a JSON(JavaScript Object Notation) file and fed into the NLP algorithm which would keywords that can group emails together and notify relevant senders and recipients. The direct users would then be able to log into a web application interface using active directory credential and modify the lists of keywords for each topic, and then send URL links to the indirect users connecting them with each other on common projects they may collaborate on. The indirect users would then able to log into a web application interface via the URL using their active directory credentials to see the keywords and names of other Bayer employees on similar projects, as well as resources from the internet that may contain additional information that is useful.

1.4 Product Overview

1.4.1 Needs and Features

Note: The timeline proposed for each phase is defined as follows:
Phase 1-- POC(Proof of Concept): late-August 2018 to mid-October 2018
Phase 2-- Development and Implementation: late-October 2018 to mid-December 2018
Phase 3-- Deployment: late-December 2018 and onward

Need	Priority	Features	Planned Implementation
Keyword Identification	High	A structured output in JSON format showing a list of keywords from a large corpora of email contents	Phase 1
User Collaboration Identification	High	An output that would not only include topics/keywords but also the senders/receivers related to these topics	Phase 1
Dynamic Learning	High	A web UI that allows admins and users to add/delete keywords for topics	Phase 2
Resource Discovery	Medium	A web UI that shows keywords grouped by topics with URLs from internet resources (i.e. BUBL Information Service)	Phase 2
User Notification	Medium	A URL output that can be sent to admins and users prompting them to login onto the web application UI, sent via email, with fixed frequencies (daily/weekly)	Phase 2
Tiered Access	Medium	The login for users would be tiered, enabling admins and general users to access different UIs	Phase 2

Connection with Bayer’s email servers	Low	The backend of the web application should be able to communicate with Bayer’s internal email servers in order to extract text data	Phase 3
Integration with Bayer’s Active Directory	Low	The login for admins and users would integrate with Bayer’s existing Active Directory	Phase 3

1.4.2 Other Product Requirements

Due to Bayer’s data security practices the project team is not able to access Bayer’s internal email data and knowledge base for the POC and Implementation Phase. Instead, the project team proposes to focus more on knowledge discovery and knowledge sharing. Academic paper abstracts in various fields are scraped from the internet and are formatted in ways that resemble an email structure. In the final Implementation Phase, Bayer META should be able to connect with Bayer’s email servers to obtain actual email data, and to extract information including but not necessarily limited to subjects, email addresses of senders and receivers, as well as email content.

1.5 Points of Contact

Client Contacts

Name	Title	Email Address
Jim Koob	Enterprise Architect, Business Intelligence and Analytics	jim.koob@bayer.com

Michael Kremliovsky	Director, Medical Devices & eHealth	michael.kremliovsky@bayer.com

Project Team Contacts

Name	Title	Email Address
Sohel Sarwar	Project Advisor	ssarwar@andrew.cmu.edu
Maggie Lu	Project Manager	yaol4@andrew.cmu.edu
Siddharth Menon	System Administrator	ssmenon@andrew.cmu.edu
Olivia Zheng	Process Manager	xinpengz@andrew.cmu.edu
Yiling Zhong	Finance Manager	yzhong1@andrew.cmu.edu

2. System Overview

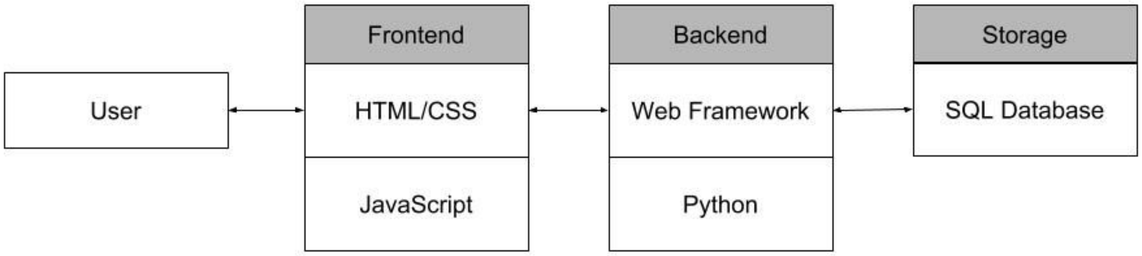
2.1 High Level Description

META aims to examine a large aggregate of text data and extract useful insights from it. These insights include: keywords, topics, authors and relevant senders/receivers of relevant emails. It needs to also be able to take user input and allow the administrators to remove certain keywords when privacy or security concerns arise. In addition, META will also be able to take user input into account when recompute its natural language processing model so the results can be more personalized and accurate.

The project team envisions META to be a web application, hosted first remotely on heroku.com during the Proof of Concept phase. The reason being that it will facilitate faster and more convenient pilot inside Bayer's own web environment. Later META will move to Bayer's internal environment where it can directly communicate with Bayer's email server to retrieve the corpora of email content in plain text.

In final implementation, META will also provide integration with Bayer's Active Directory for user credentials. There are mainly two tiers of users for the system as of now: the administrators will have the ability to oversee and omit keywords/topics as they deem fit; the users will be notified via emails that contain URL links. After logging in, the users will be able to see lists of relevant topics that can be helpful to specific ongoing projects, along with other users who may be interested in the same projects, and internet resources (see section 5.2 User Interface for details).

2.2 Technology Stack



3. Technical Approach

3.1 Tools

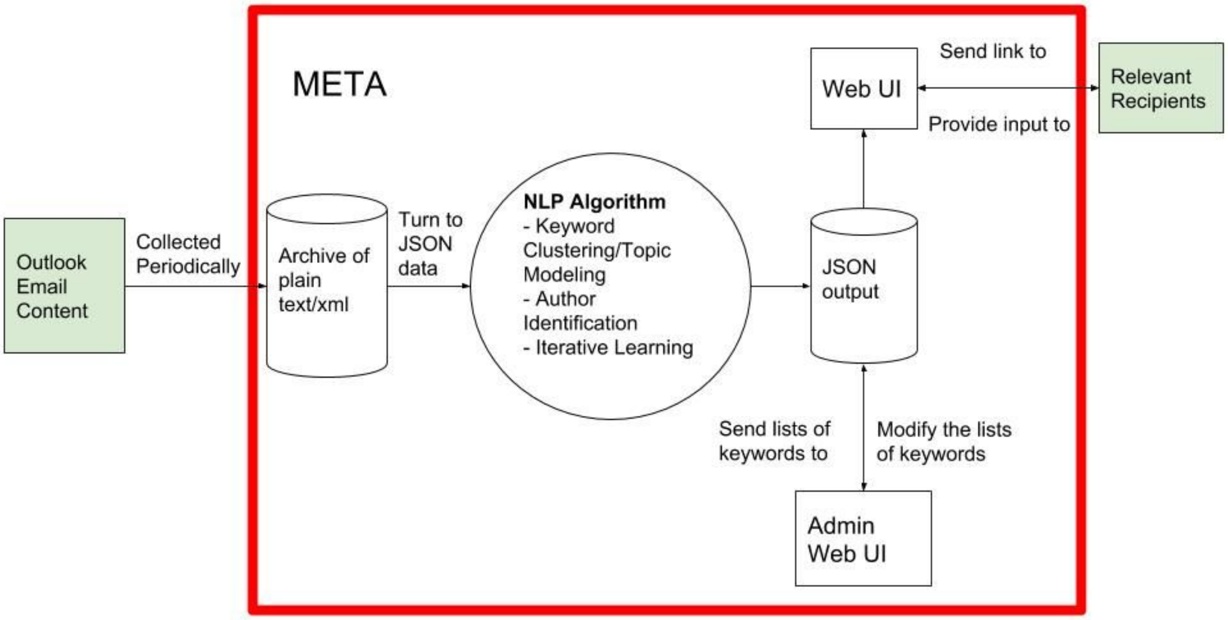
Jupyter Notebook: Jupyter Notebook is an open source web application that contains live code in multiple languages, including Python, which is the backend language choice due to its readability and its robust collection of third-party open source libraries for natural language processing. In the POC phase, all the data cleaning, processing and analytics are done in Jupyter Notebook for quick demo purposes. The python scripts can be easily repurposed later for implementation purposes.

Gensim Library: Gensim is a robust open source vector-space modeling and topic modeling toolkit implemented in Python. Gensim is specifically designed to handle large text collections, using data streaming and efficient incremental algorithms, which differentiates it from most other scientific software packages that only target batch and in-memory processing.

Web frameworks: the project team has investigated and proposes two frameworks as potentially feasible solutions. Both Flask and Django are well-known Python enabled frameworks, while Flask being more customizable and low-level focused. The project team also recommends using Heroku as the web application host due to the team's familiarity with its comprehensive PaaS (Platform as a Service) offering. The team may also use the AJAX libraries within jQuery or AngularJS in the future to handle the RESTful requests that the frontend users would be making to communicate with the NLP algorithm on the backend.

4. System Architecture

4.1 High Level Architecture



Ideally, META collects data from Bayer's outlook email servers and turns it into structured JSON file. The JSON file will then be fed into the NLP algorithm where the keywords, topics and relevant senders/recipients of emails will be discovered and identified. The result will then be formatted into a JSON object and shown on the user interface of the web application accessible by users.

4.2 Deployment

META can be broken down into two major components: the user interfaces and the NLP algorithm. In order to deploy the UI elements, the project team will be using Heroku as the host server. The UI would act as dashboards, showing the user or the administrator relevant information based on the output provided by the NLP algorithm.

The second key element is the NLP algorithm itself. The algorithm is currently housed as a local application that is run on a local computer within the Jupyter Notebook environment. Going forward, the project team plan to move the application into the web in the form of a web application. This web application would act as an API endpoint that would be accessed using REST services. Taking a file as input, the application would analyse the data and provide the relevant information back in the JSON format. This data would then be processed by the user interfaces and displayed accordingly.

4.3 Application Pipeline Design

Web UI: The Admin UI is being created using HTML5, CSS3, Bootstrap and JavaScript. An administrator will be able to log into the UI and look at the different keywords being used in the organization. The administrator would also be able to blacklist certain keywords. These keywords would no longer be used in the calculation of topics and keywords.

NLP Algorithm: Currently the NLP algorithm is being run locally through Jupyter Notebook. The project team will deploy the algorithm on the web so that it would serve as an API endpoint for the application. META would send the collected email data directly to the algorithm which would process it and store a response JSON as an intermediary step.

5. Detailed Design

5.1 Analytical Methodology

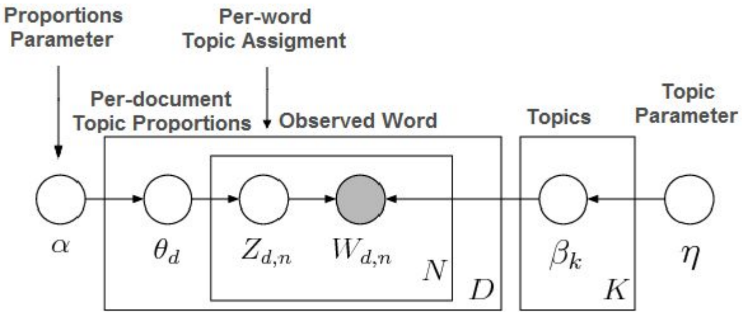
5.1.1 Data Collection and Formatting

Due to the client's concern about data privacy and safety, the project team is not able to obtain client data for this project. As a workaround for data sources for the NLP algorithms, the project team scraped various academic journal websites and formatted the content into pseudo-email forms (see Appendix 1) stored as a JSON file, per our client's request and preference. The sources include Journal of the American Chemical Society and Europe PubMed Central, an open repository of biomedical and life sciences research papers.

The content of these entries main focus on various fields within the life sciences discipline (chemical engineering, biological engineering, medical practices, etc). The project team not only wants the data for the POC phase to be an accurate simulation of email content, but also a large corpora of texts with interconnected topic and subject matters in order to demonstrate the effectiveness of the NLP algorithms.

5.1.2 LDA vs TF IDF vs LDA_Mallet

In natural language processing, latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. In our context, LDA can be applied to extract topics from one or more documents by grouping similar words together.



TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. TF IDF helps us to filter out common but not important adjective or verb in the documents, and therefore gives the model a clearer result.

$$\mathrm{TF}-\mathrm{IDF}=\mathrm{tf}(t_i,q)\times IDF(t_i)$$

$$TF(t,d)=freq(t,d)$$

$$IDF(t,D)=\ log_2\frac{|D|}{|\{d\in D|t\in d\}|}$$

Developed by University of Massachusetts, MALLET topic model package includes an extremely fast and highly scalable implementation of Gibbs sampling, efficient methods for document-topic hyperparameter optimization, and tools for inferring topics for new documents given trained models. This option turns on hyperparameter optimization, which allows the model to better fit the data by allowing some topics to be more prominent than others. Optimization every 10 iterations is reasonable.

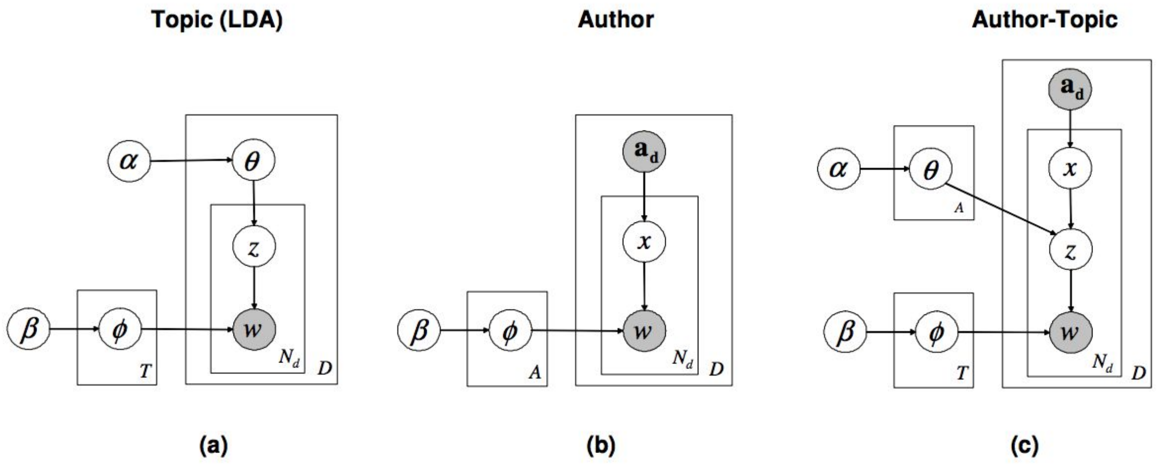
To validate the models and choose the optimal numbers of topics, we adopt the coherence score as the metric. The state-of-the-art in terms of topic coherence are the intrinsic measure UMass and the extrinsic measure UCI, both based on the same high level idea. Both measure compute the sum of pairwise scores on the words w_1w1, ..., $w_n$wn used to describe the topic, usually the top nn words by frequency p(w|k)p(w|k). The higher the coherence score, the better the model is.

$$\text{Coherence}=\sum_{i<j}\text{score}(w_i,w_j)$$

By looking at the coherence score results from iterations, we will take the point from which the score starts to flatten out as our number of topics. Since there could be more than one such point, we will pick the one that best fits the expected outcome.

5.1.3 Author-Topic Modeling

Author-Topic Modeling is a direct extension of Latent Dirichlet Allocation (LDA) topic modeling. In addition to model topics, it also includes authorship information and is primarily used within a large corpora of text for author matching (see Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (n.d.). *The Author-Topic Model for Authors and Documents*). We find this method to be applicable also in identifying email senders/receivers given a body of email texts.



There are two major components in the Author-Topic Modeling, the first being document topic modeling, and the second being author association with keywords. In respect to document topic modeling, LDA is a commonly used algorithm (see above section for more details) and models documents into mixtures of probability distributions. As shown in figure (a), φ denotes the matrix of topic distributions, with a multinomial distribution over V vocabulary items for each of T topics being drawn independently from a symmetric Dirichlet(β) prior. θ is the matrix of document-specific mixture weights for these T topics, each being drawn independently from a symmetric Dirichlet(α) prior. For each word, z denotes the topic responsible for generating that word, drawn from the θ distribution for that document, and w is the word itself, drawn from the topic distribution φ corresponding to z. Estimating φ and θ provides information about the topics that participate in a corpus and the weights of those topics in each document respectively.

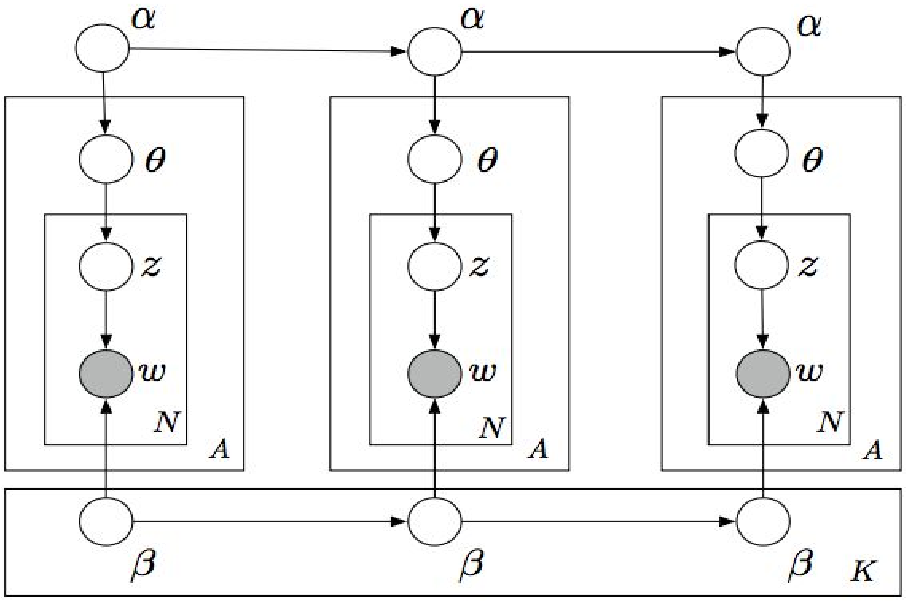
Although the topic modeling model gives us an interpretable model for topic identification with keywords associated with each topic, it does not show anything about the information of the author. Therefore we need to incorporate the author model. For each word in the document, an author is chosen uniformly at random, and a word is chosen from a probability distribution over words that is specific to that author. As shown in figure (b), x indicates the author of a given word, chosen uniformly from the set of authors ad. Each author is associated with a probability distribution over words φ, generated from a symmetric Dirichlet(β) prior. Estimating φ provides information about the interests of authors, and can be used to answer queries about author similarity and authors who write on subjects similar to an observed document.

The Author-Topic Model draws the strength of both models. As in the author model, a group of authors decide to write the document. For each word in the document an author is chosen uniformly at random. Then, as in the topic model, a topic is chosen from a distribution over topics specific to that author, and the word is generated from the chosen topic.

5.1.4 Dynamic Learning

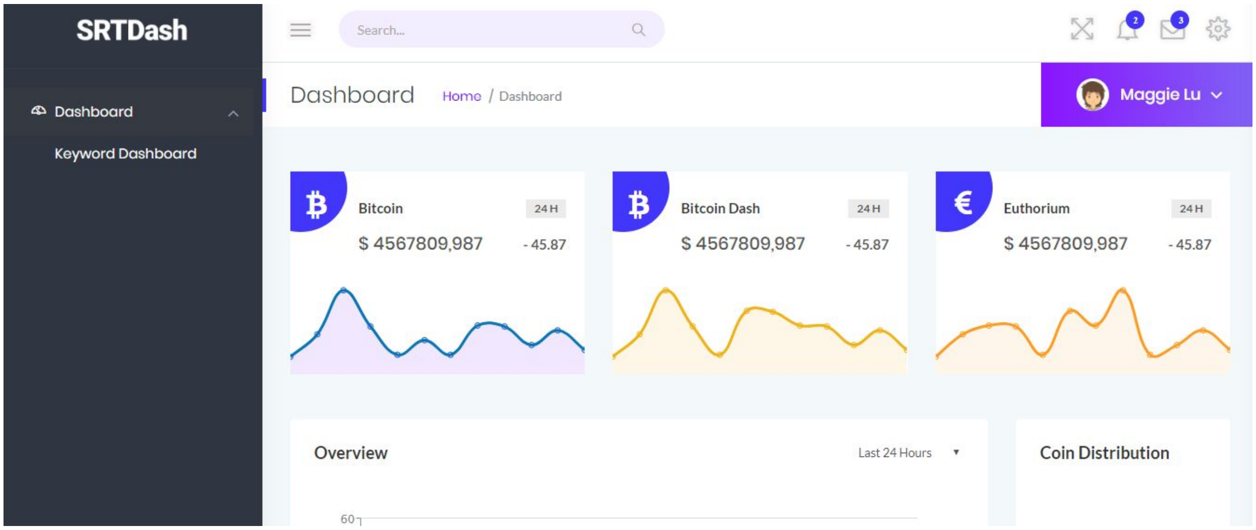
Dynamic Topic Modelling is an extension of Latent Dirichlet Allocation (LDA) topic modelling to handle sequential documents. In a dynamic topic model, each document is viewed as a mixture of unobserved topics while each topic defines a multinomial distribution over a set of keywords. Documents are grouped by time slice and it is assumed that the documents of each group come from a set of topics that evolved from the set of the previous slice. For each word of each document, a topic is drawn from the mixture and a keyword is subsequently drawn from the multinomial distribution corresponding to that topic. Thus while the topics remain the same, their keywords evolve over time slices and paint different pictures of the text. An example is shown in Appendix 2.

The figure below is a graphical representation of a dynamic topic model for three time slices. Each topic's natural parameters βt,k evolve over time, together with αt, a parameter for the topic proportions.



5.2 User Interface

The user interface has been developed using Bootstrap to enhance the initial interface built using HTML and CSS. A mockup of the initial interface is as shown below:



Currently the data being shown is static data entered manually. The interface would be similar for the web and admin interfaces, the only difference being the actions that an administrator would be able to perform (including blacklisting a keyword and other administrator specific activities). These interfaces would be updated based on the JSON output that we receive from the NLP algorithm.

5.3 System Integration

For front-end integration: the project team envisions META to be a stand-alone web application, therefore the users should be able to access META via the major browsers (Firefox, Chrome, etc). META will also integrate with Bayer's Active Directory system to provide user login authentication.

For back-end integration: once META gets moved into Bayer's corporate network, META will make encrypted API calls to Bayer's Microsoft Outlook email servers to retrieve email content on a nightly basis, transforming and storing the data as JSON file on the server META files reside.

6. Appendix

Appendix 1: Sample Email Format (For Proof of Concept)

From the dataset:

{

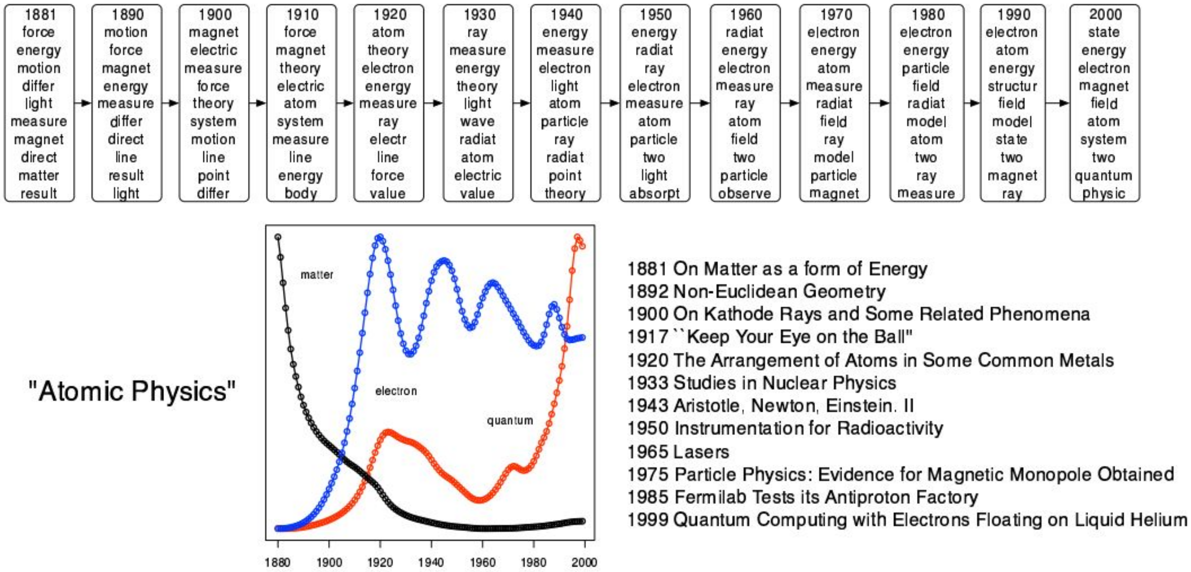
"Title": "Data on biosurfactant assisted removal of TNT from contaminated soil.",

"Sender": "Bijan Bina",

"Receiver": "Mohammad Mehdi Amin",

"Content": "Contamination of environment, especially soil, is in great concern and can cause health problems. Thus, remediation of these pollutants through environmentally friendly methods should be considered. The aim of this data was bioremediation of TNT from contaminated soil. Two plastic pans were used as bioreactor. In each pan, 3/202/kg of soil was used. Concentration of TNT in contaminated soil was 1000/202fmg/kg. Rhamnolipid in concentration of 6002mg/l was added to intended pan. Sampling was done in each two weeks. In order to assessment of TNT degradation, samples were analyzed with HPLC. The data showed that after 154 days of experiment, TNT removal in soil that amended with rhamnolipid was 73% and in experiment with no addition of rhamnolipid was 58%. Based on the obtained data rhamnolipid was effective in remediation of TNT contaminated soil."}

Appendix 2: Example of Dynamic Topic Modelling



1881	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
force	motion	magnet	force	atom	ray	energy	energy	radiat	electron	electron	electron	state
energy	force	electric	magnet	theory	measure	measure	radiat	energy	energy	energy	atom	energy
motion	magnet	measure	theory	electron	energy	electron	ray	electron	atom	particle	energy	electron
differ	energy	force	electric	energy	theory	light	electron	measure	measure	field	structur	magnet
light	measure	theory	atom	measure	light	atom	measure	ray	radiat	radiat	field	field
measure	differ	system	system	ray	wave	particle	atom	atom	field	model	model	atom
magnet	direct	motion	measure	electr	radiat	ray	particle	field	ray	atom	state	system
direct	line	line	line	line	atom	radiat	two	two	model	two	two	two
matter	result	point	energy	force	electric	point	light	particle	particle	ray	magnet	quantum
result	light	differ	body	value	value	theory	absorpt	observe	magnet	measure	ray	physic

1881 On Matter as a form of Energy
1892 Non-Euclidean Geometry
1900 On Kathode Rays and Some Related Phenomena
1917 ``Keep Your Eye on the Ball"
1920 The Arrangement of Atoms in Some Common Metals
1933 Studies in Nuclear Physics
1943 Aristotle, Newton, Einstein. II
1950 Instrumentation for Radioactivity
1965 Lasers
1975 Particle Physics: Evidence for Magnetic Monopole Obtained
1985 Fermilab Tests its Antiproton Factory
1999 Quantum Computing with Electrons Floating on Liquid Helium

This graph from *Dynamic Topic Models* (Beil and Lafferty 2006) shows how the topic of Atomic Physics involves different keywords over time. Using years as time slice, the model shows how research in this field evolved over time to give the topic a new meaning.