Projet M1 UFR MIM : Sujet IR Taxinomie MISP

Marwin NIMESKERN 1
Grégory GUGGENBUHL 2
Jules CHARRON 3

1 Université de Lorraine
UFR MIM
3 Rue Augustin Fresnel, 57070 Metz
marwin.nimeskern5@etu.univ-lorraine.fr
2 Université de Lorraine
UFR MIM
3 Rue Augustin Fresnel, 57070 Metz
gregory.guggenbuhl7@etu.univ-lorraine.fr
3 Université de Lorraine
UFR MIM
3 Rue Augustin Fresnel, 57070 Metz
jules.charron5@etu.univ-lorraine.fr

Taxonomie MISP DARWIN

M. NIMESKERN, G. GUGGENBUHL et J. CHARRON

Francine HERRMANN

Gratefully

We are tankful to the MISP team who despite a lack of time have open their door to make some reunion and to present us the MISP's software.

Contents

Introduction

What is a Taxonomy ?

Nowadays, many scientist project are founded on classification more or less elaborate.

We will talk about : taxonomy

was created by the Swiss botanist Augustin Pyrame de Candolle in 1813, for a species classification in herborize purpose.

Over time, this term was extended to other areas of science, including the humanities, information sciences or computer science.

The taxonomy designs the practice and science of classification of things or concepts, including the principles that underlie such classification.

A Taxonomy is a classification of information instances that facilitates the discovery and management of information between them.

Goal of our Taxonomy

Misp (see next page for further explanation) is grouping all security incidents that happened in all enterprises which would share information with other enterprises. This information could be important and concerns different level of an enterprise.

The main questions

- What are these information and how could it be important for each work of an enterprise?
- How to classify all this information in order to sort the most important?

Our goal is the creation of a cooperative taxonomy, in order to classify all the instances of enterprises. This will allow to make a better search and classification of all the events of MISP.

What is a Corporate taxonomy?

A corporate taxonomy is the hierarchical classification of entities of an enterprise, organization or administration, used to classify documents, digital assets and other information. Taxonomies can cover virtually any type of physical or conceptual entities (products, processes, knowledge fields, human groups, etc.) at any level of granularity.

State of Art

MISP Malware Information Sharing Platform and Threat

Presentation

MISP is a Luxembourg government organization that brings together a community of more than 6,000 companies sharing between them information about attacks, frauds, ...

MISP is an open source software on GitHub, allowing the sharing of attacks revealed by the community between MISP instances.

Main Advantages

- One of the largest attack databases in the world
- Internationally recognized
- There are multiple data
- Taxonomies can be used as object typing
- Open-Source and regularly maintained by a warm community.
- Can look like SQL (Entity / Relationship)
- Has a partnership with the University of Lorraine

Main Drawback

- Unusual in the foreground, requires a good understanding of the software and its modules
- little concrete application on the data
- Only coded in python that a really slow language because it's a scripted language

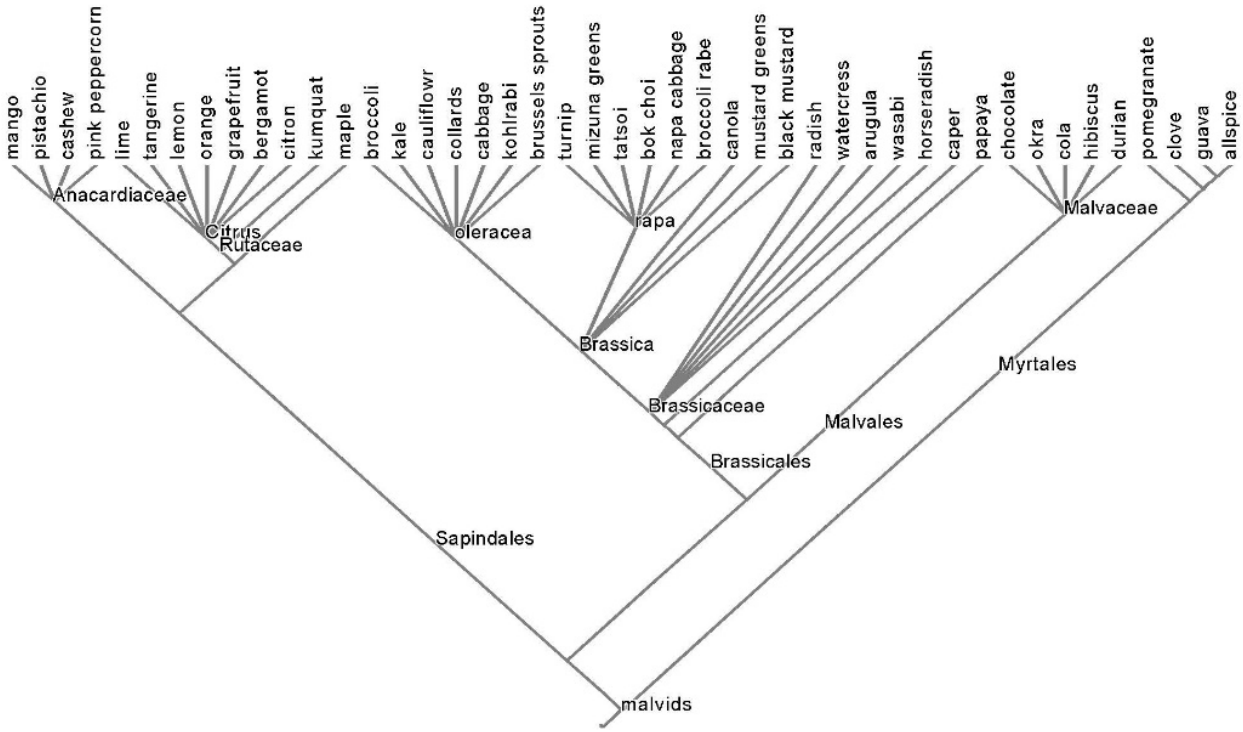
Conclusion

The attributes are interconnected and juxtaposed with a generic taxonomy: this will allow typing and classifying all events generated in MISP

The arborescent Taxonomy

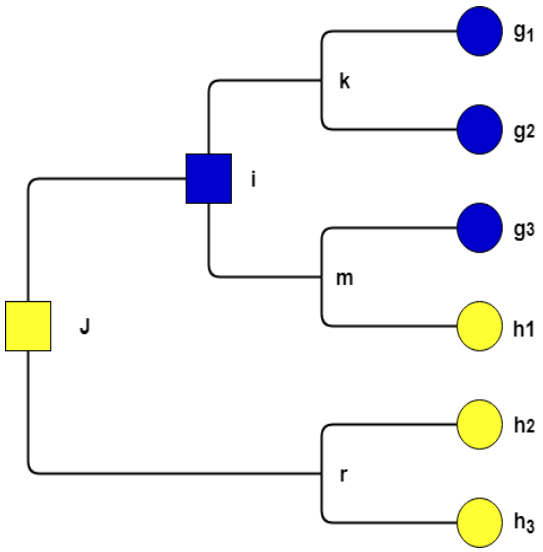
Serving as the basis of scientific classification of species, the term "Taxonomy" refers in computer science to a method of classifying information in a structured architecture so that it can be completed in an evolutionary and generic manner.

Example of a tree of a botanical taxonomy:



Phylogenetic tree view Tree taxonomies are kinds of graphs that help you to better classify an object into several families of class and subclass. These kind of taxonomy have oriented our research to choose our corporate taxonomy.

A tree taxonomy is presented by this generic form



The MISP

The MISP's team have of course create an engine to manage their taxonomies.
Therefore, we have adapted our taxonomy to best fit their system

Presentation

The taxonomies of MISP are in the form:

namespace.predicate = "value"

- Namespace : Name of the taxonomy
- Predicate : Object resource of the taxonomy
- • Value : Value of the resource
- : resource taxonom

As a result of these rules, we make the json in annexe 2 this annexe is here to describe how a generical taxonomy is written in order to be implemented directly in the MISP project.

Main Advantages

Misp use taxonomies engine to create statistics, rate and classify events

Main Drawbacks

- Without predicate parsing, the depth is limited to 1.

  To avoid these drawback:

  - A predicate could be parse to create multiple levels of tags and create a deeper taxonomy. Due to our need to depth, our parser will be the: _
    - Predicate: predicate = ''value"
- Long taxonomy can be problematic to use

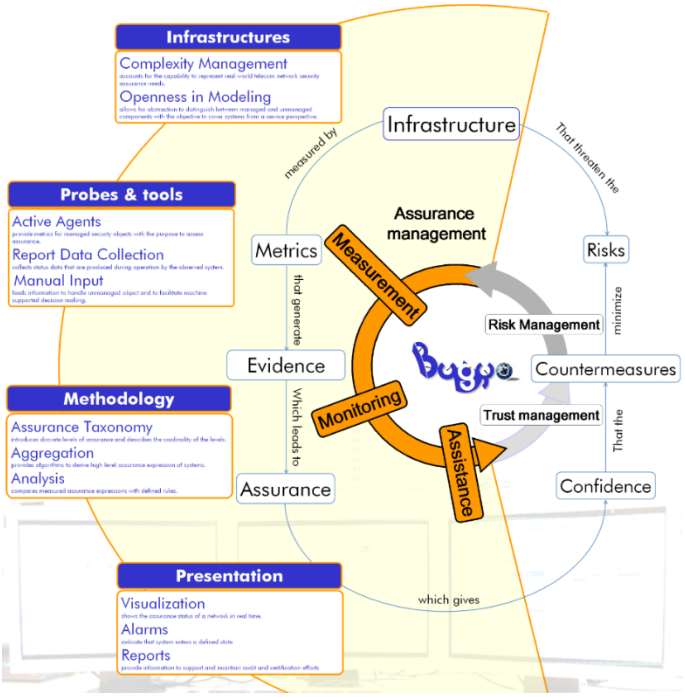
  To avoid these drawback:

  - A long taxonomy should have an interface for proper use
  - A complex need taxonomy need structuration to be used properly that why we will be inspired by other taxonomy in order to achieve that goal.

Bugyuo Methodology

Presentation

Security Methodology Project for Telecom Operators



The infrastructure is generated by metrics that generate evidence that serves to minimize the risks that give a level of confidence on countermeasures that minimize risk.

See the modification annexe to see the bugyo dendrogram use for this taxonomy.

Main Advantages

- Gives an interesting approach on the methodology of a security policy, security assurance and their different implementation and operation.
- Gives an interesting approach on the methodology of a security policy, security assurance and their different implementation and operation.
- Created with the collaboration of one of our security teachers who come from Luxembourg.
- Parallel approach to common criteria and deploys an interesting risks approach.

Main Drawbacks

- Abstract concept which only give indications on how to proceed an engineer to securing a system
- The notions are very high in the methodological approach.
- Even though this representation is powerful, it is unknown to the general public

Conclusion

To better understand the needs of an ISS, this methodology can help us to see if an event responding to its needs. These is a these criteria have none really main methodology or program to applied them. In that goal, a methodology like bugyo could be to check the advanced of these methodology in their enterprise. Many security events could be marked to follow the proper application of these stuff, furthermore a sharing could be to see how properly manage it. The more the number of companies using it and sharing their ways of proceeding will increase, the more it will be applied properly and understandable.

ERCOT Faceted Classification

Function	Activity	Type	Entity Type	Entity	Rule Type	System Name
Market: Participation	Registration and Qualification	Registration Documents	Market Participant		Protocol	
Information Technology	System and Application Development	Revision and Change Request	Market Participant	TDSP		MarkeTrak

Predicate	$\perp$	Predicate
Value	$\perp$	Value
Value	$\perp$	Value

Lexique : A $\perp$ B : A independent of B
Statically describes the operation of a business and its activities.
Each column describes a predicate of an instance from the service of the company for a proper purpose

Main Advantages

- Briefly presents the operation of a department of a company
- Hardcoded approach for corporate taxonomy
- Representation which strongly resembling a multi-model-level (further details in this document)

Main Drawbacks

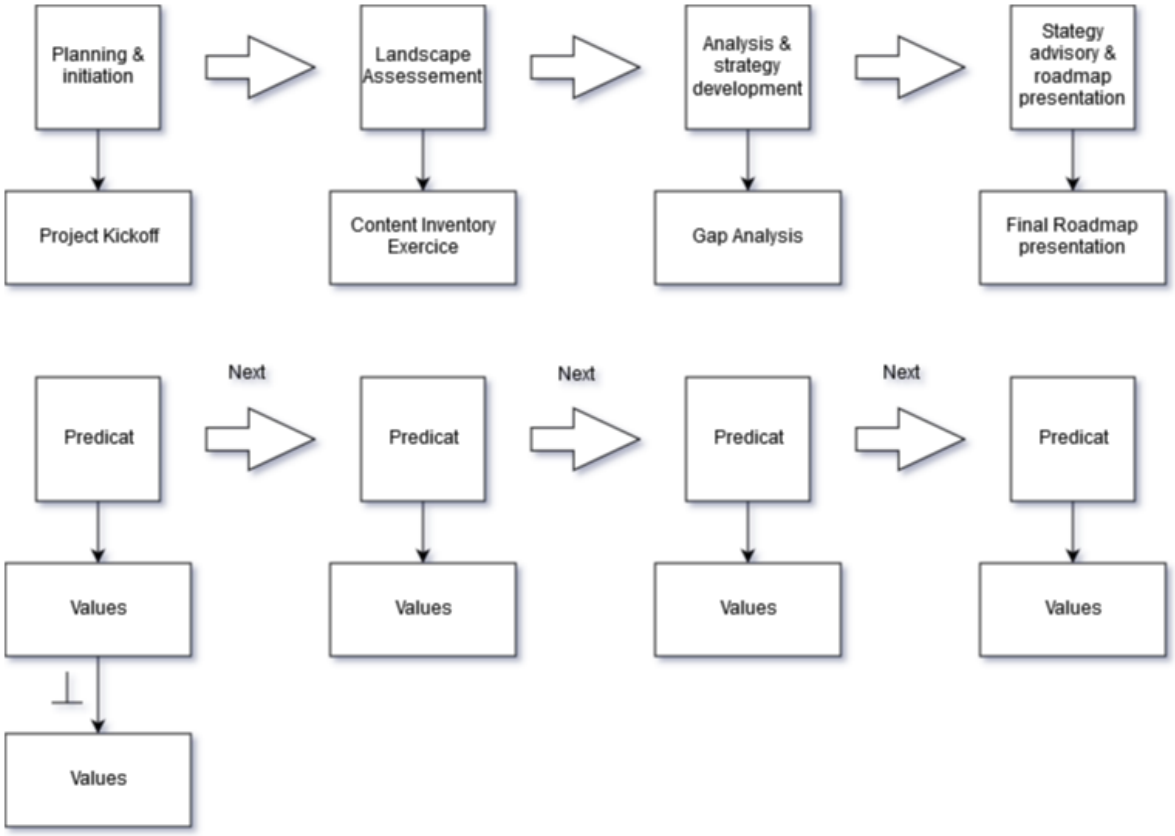
- Too simple to use alone without another taxonomy
- All links of columns are independent from each other, it a drawback only if no other taxonomy is used to described how to fill the column

Conclusion

- Despite these drawbacks these taxonomies could be used as a model for a properly structuration for another taxonomy like the one we decided to choose.
- It could have inspired a way to present an object created with our taxonomy

  As announced, the columns is only a drawback if no other taxonomy is used, more over it became a strongly advantage in the goal of describe a multi-level-model.

Architecting an Enterprise content management Strategy



Present the strategies of a company for the realization of a project.
The arrow presents the next instance that will be needed to realise the project.
Each predicate could have many values, which will be independent between them.

Main Advantages

- Present a taxonomy, in a timeline manner
- Could be useful to describe instance or event between them in a chronological manner

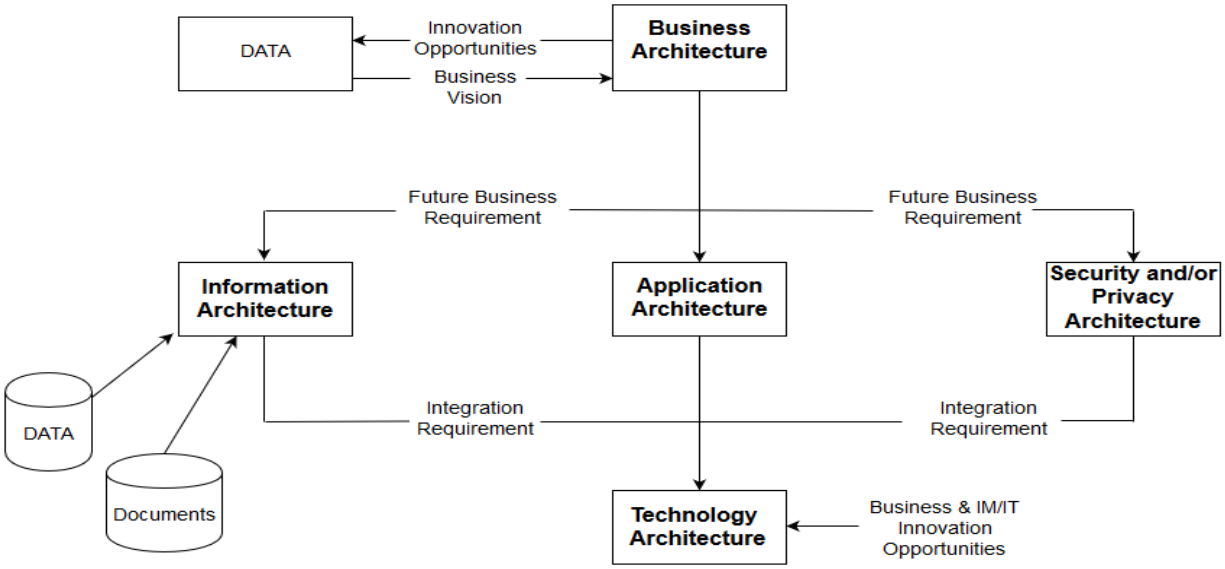
Main Drawbacks

- Does not in any way present the enterprise infrastructure, it's useless for present jobs or objects

Conclusion

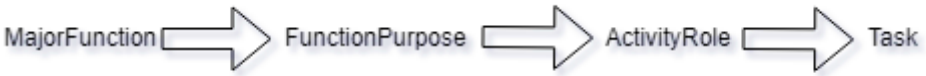
- Our job is not only to describe the infrastructure of an enterprise but all instance of an enterprise therefore it's could be important to make some way to describe a chronological order for some plans of a company
- It could be an interesting way to structure a task in order to accomplish a company project.

Functional Classification Taxonomies



This graphics explain how for them a data is share in some architecture:

- Business: harvest the data information
- Information: store the data in some disks
- Application: Use this data for the purpose of the company
- Security / privacy Secure the data harvested
- Technology: has been used by the application part to process the data



Business / Application / Technical / Information
Define links between the major instances of an archetype company.
Education => Teach => Teacher => Network course

Main Advantages

- The major instances of a company could be linked thanks to that

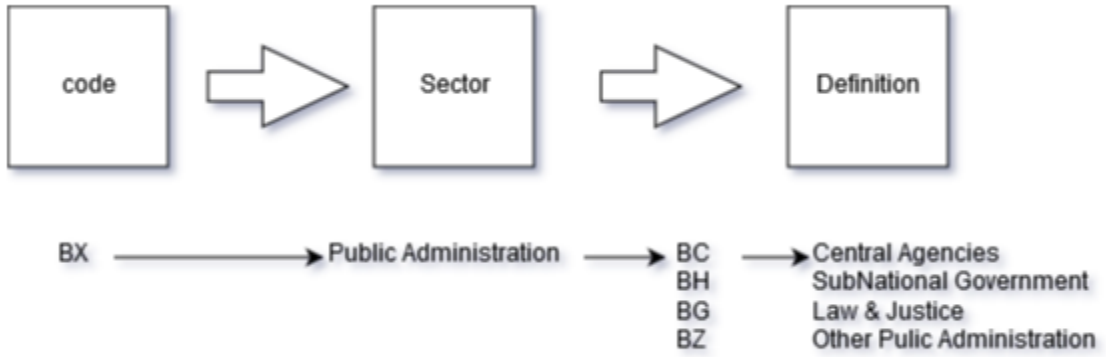
Main Drawbacks

- Doesn't explain what is an instance of an enterprise

Conclusion

- The main idea of linking could be used to linked other instances from a taxonomy more complex like the GEA-NZ taxonomy

Sector Taxonomy and definitions



Interesting mapping of instances of jobs from any sector of activity

Main Advantages

- Professional directory of classification
- Easy way to research an instance once the code is know

Main Drawbacks

- More a directory of job than a real taxonomy

Conclusion

- It could be interesting, once our taxonomy operational to map our instance realised thanks to that way and a hierarchical methods

Taxonomy code mapping / professional providers

It's a directory of instance, consequently it's presented the exact same advantages and drawbacks than the previous Sector Taxonomy and definitions.

ICB Version 2: IPMA Competence Baseline

Example of ICB-Version page 74

7	loyalty, solidarity, readiness for helping Loyalität, Solidarität, Hilfsbereitschaft loyauté, solidarité, aide				
Nr.	**Characteristics, Merkmal, Caractéristiques**	**+**	**0**	**-**	**Opposite, Gegensatz, Opposition**
7-1	accepts the rules on team co-operation, supports team decisions akzeptiert Spielregeln der Kooperation im Team, unterstützt Gruppenentscheidungen accepte les règles de coopération dans l'équipe, défend les décisions de l'équipe				ignores agreed rules, does not accept team decisions consequently hält sich nicht an abgemachte Spielregeln, akzeptiert Teamentscheidung nicht unbedingt ignore les règles convenues, n'accepte pas toujours les décisions de l'équipe
7-2	defends the team against outside, if necessary, is loyal to team members verteidigt das Team nach außen wenn nötig, ist loyal zu Teammitgliedern défend l'équipe à l'extérieur, si nécessaire, est loyal aux autres membres de l'équipe				is reluctant to outside, discloses confidential team information to outside hält sich nach außen zurück, bringt Vertraulichkeiten nach außen renâcle à défendre l'équipe à l'extérieur, révèle à l'extérieur des informations confidentielles sur l'équipe

A taxonomy asking a company of their competence.
Present a really interesting way to make a survey in order to create an instance from this taxonomy.

Main Advantages

- The survey is a really interesting and easy way to asking a thing to a customer

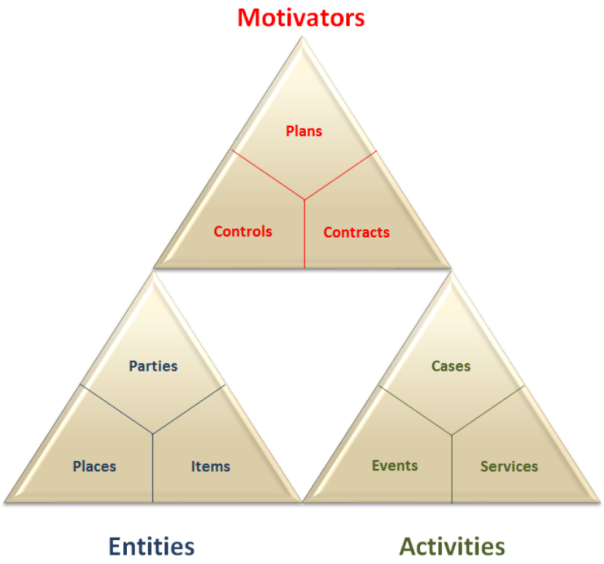
Main Drawbacks

- The taxonomy itself isn't really interesting to make a corporate taxonomy

Conclusion

- Once our taxonomy is designed we could use this way to ask for create dynamically some instance of our taxonomy.

GEA-NZ: Government Enterprise Architecture New-Zealand



The taxonomy that has been choose by us.
Thanks to three complete tables also divided in three branches
Note : See the annexe 2 to see the main branch presented without their leaf for more readability See the NGA-NZ for properly instructions about their leaf, the taxonomy being used as it is been , the rewrite here would have no interest except for the add we made of it.

Description of the main branch

- **Motivator:** Purpose
  - **Plans:**actions accomplished for choose a direction
  - **Controls:**the constraints of actions
  - **Contracts:** agreement between different parties
- **Entities:** all possibly instances in enterprise
  - **Places:**Workplace and any location
  - **Items:**All items buyable or not
  - **Parties:**Staff entities
- **Activities:**all possibly interaction between parties
  - **Cases:**Special and normal event managed
  - **Events:**Planning and / or Spontaneous event
  - **Services:**Offer of the service

Main Advantages

- A really complete cooperative taxonomy for each instances of an enterprise

Main Drawbacks

- It's as complete as it is complex to use, that's why It's need a structuration for a properly usage
- Some names could be put in a generic: new-zealand could be replaced by enterprise headquarter

Conclusion

This taxonomy is powerful but it does not force any structuring to make some object like : jobs, activities.
This structure is crucial to make some classification furthermore.

Modification of GEA-NZ

Note: Please see the Annexe 3 for the picture of adaptation into GEA-NZ taxonomy.
In purpose to make the control region more accurate we add two more sub-category.

Risk Governance

- Risk Governance in order to evaluate the risk taken with the object
    - **Residual Risk :** This element is a remaining risk in the enterprise
    - **Risk Analysis :** Risk by the enterprise
    - **Risk Assessment :** General process of analysis and risk assessment
    - **Risk Management :** Control and organize the activity taking into account the risks
    - **Risk Treatment :** Measures selections and implements the chosen metrics

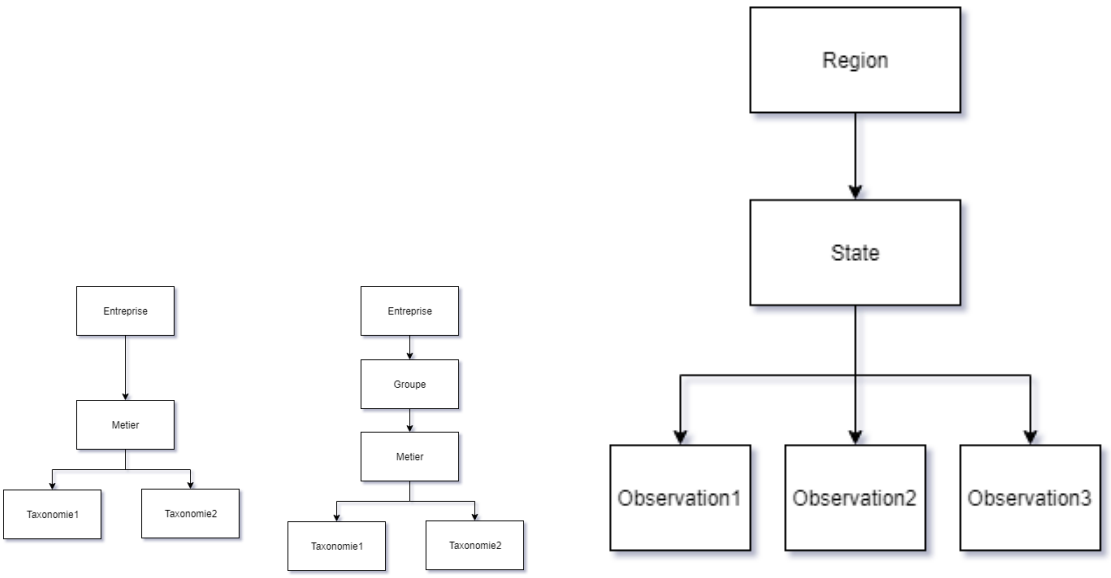
Bugyo

- Bugyo in order to add some security methodology in the code to verify directly a component
- **Service model:** applied to the model as a whole
    - **Absence of relevant vulnerability**
        - Is a vulnerability present? yes // no?
    - **Unmanaged/Managed Ratio**
        - Less Unmanaged / Higher confidence about the ratio ?
- **Maintenance Management** applied to the model to see the maintenance of it
    - **Probe Maintenance:** Is the maintenance
        - **Reactive :** Ad-hoc maintenance / maintenance policy present ?
        - **Pro-active:** Regular interval to prevent / Documentation and log
    - **Infrastructure object model maintenance:** maintenance for each object
        - **Reactive :** Ad-hoc maintenance / maintenance policy present?
        - **Pro-active:** Regular interval to prevent / Documentation and log

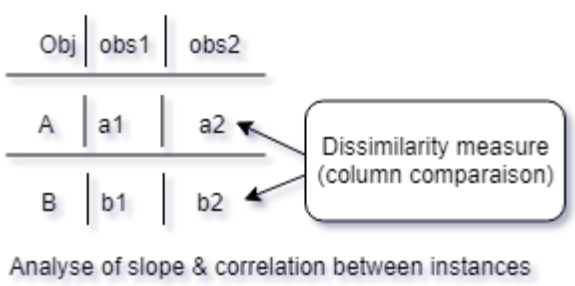
- **Metric Construction:** is used to measure the system's security policy
    - **Scope:**request the number of elements measured by the system
        - Are there countermeasure? How much it's goes around the system?
    - **Depth:** How for in details it's goes?
        - Is there an interface for?
            - † o A single object? The whole chain?
              The whole chain and all the values passing through it.
    - **Rigor:** How is the formalism of the measures? Is it strict?
        - None verification / informal-verification / semiformal verification
    - **Reliability of metric** Are the measure repeatable?
        - None verification / reproductible / informally / semiformal
    - **Timeliness:** How recent are the software?
        - None verification / Version prior the newest (if retro compatible)
          Most recent version (Avoiding flaws with patch)
    - **Frequency:** How fresh is the evidence of measure?
        - Year / Month / Week / Day / Hours
    - **Stability:** How stable is the measure
        - Is there a human avoiding False/True positive?
          Or / And are there statistics algorithms?

The multi-level-model (MLM)

Multilevel models (also known as hierarchical linear models, nested data models, mixed models, random coefficient, random-effects models, random parameter models, or split-plot designs) are statistical models of parameters that vary at more than one level.
A multi-level-model is created when a complex structuration is used and require different cutting levels.
Each region could have many states and each state could have many observations and so on.

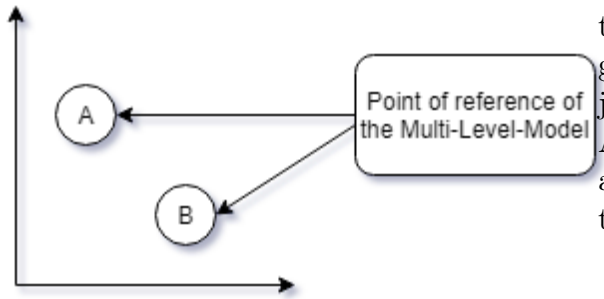


Generical Taxonomy ***enterprise job*** ***enterprise service of job***



Analyse of slope & correlation between instances

For each object created we have a level and for of the could be used
In order to make possible this stuff we could make for each type of object a table in order to compare an object between another object of the same type. (like job, service)
Once a table is made for an object, we can observe a dissimilarity measure, reporting differences between the different columns.



In the end, we can make a graphics thanks to the dissimilarity measure to make a graphics of point of reference for each object.
As a result, with these graphics we can make an HAC (Hierarchical Ascending Classification

Hierarchical Ascending Classification (HAC)

In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. An HAC is created

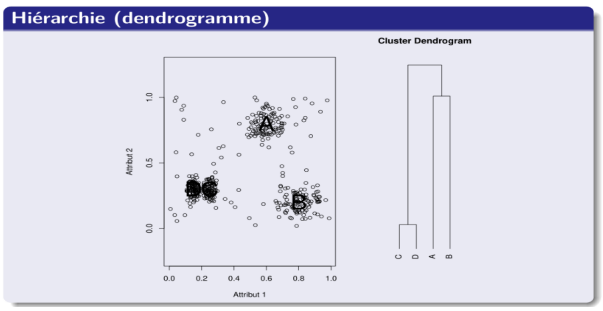
- a is needed
- **dissimilarity measure**
- for clustering many in order to make cluster analysis and segmentation.

Once created, this cluster will make it possible to present wiser choices to the client.
Point

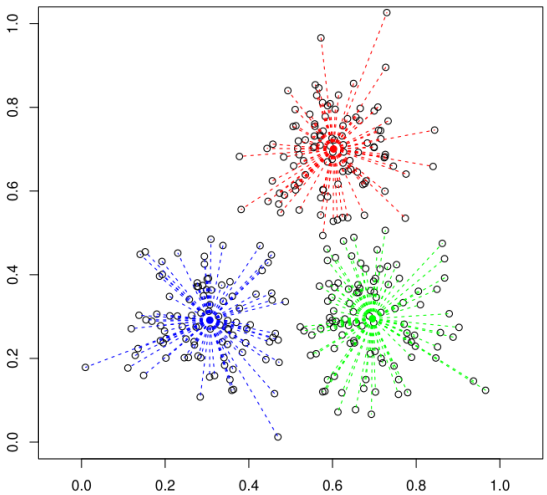
- As see each point will present an object created with the taxonomy

Strategies for hierarchical clustering generally fall into two

- **Agglomerative:** This is a "**bottom-up**" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive:** This is a "**top-down**" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.



These algorithms are usually based on distance (prototype or density) and / or Grid-based like Dendrogram (discretization of data)
These methods can make their own concept thanks to cloud of points
The cluster dendogram will be the final goal of this project



This picture represent a method based on prototypes

- A prototype is:
  - A representative individual
  - A seed (called an Isobarycenter*)

The center of each class is also called an isobarycenter
Finally, once this step arrived we must minimised by calculating the intra class inertia and then deducing each class for each object and their relation among them in order to finally made the final cluster dendrogram.

Algorithms K-MEANS

The K-means method is an optimization algorithm of intra-class inertia

Algorithms Names

- K-means, K-moyennes, K-noyaux
- Algorithm of dynamic clouds, mobile centers
- Lloyd's algorithm

The algorithm K-MEANS to determine new class and there Isobarycenter

Description

- Alternation of two stages:
    - Grouping objects into classes around centers
    - Adjustment of centers according to the component objects classes
- Stopping the algorithm when there is no more change (or after a fixed number of iterations)
- Convergence of the algorithm in a guaranteed finite time // (usually correct result in 10 iterations)

Realization

- Random initialization of centers
- As long as the result varies
- For all objects
    - Calculate the distance to all centers
    - Assign the object to the class most close
- For all classes
    - Calculate the center of gravity of the objects assigned
    - Assign the center of gravity as new center of the class

The K-means methods will only serve to create each class, the final cluster dendrogram will be created thanks to the UPGMA methods for more precision see the annexe document to made the clustering.

UPGMA : pair group method with arithmetic mean

The UPGMA algorithm constructs a rooted tree (dendrogram) that reflects the structure present in a pairwise similarity matrix (or a dissimilarity matrix). At each step, the nearest two clusters are combined into a higher-level cluster. The distance between any two clusters A and B each of size.

Determine the matrix for UPGMA

- initialize the matrix A
    - For all objects classes
    - Calculate the distance to all others objects class
- Make the UPGMA algorithm

Algorithm

The UPGMA algorithm constructs a rooted tree (dendrogram) that reflects the structure present in a pairwise similarity matrix (or a dissimilarity matrix). At each step, the nearest two clusters are combined into a higher-level cluster. The distance between any two clusters A and B , each of size (i.e., cardinality) |A | and |B|,is taken to be the average of all distance d(x,y) between pairs of objects x in A and y in B, that is, the mean distance between elements of each cluster:

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y)$$

In other words, at each clustering step, the updated distance between the joined clusters A U B and and a new cluster X is given by the proportional averaging of the dA,X and dB,X distances:

$$d_{(\mathcal{A} \cup \mathcal{B}),X} = \frac{|\mathcal{A}| \cdot d_{\mathcal{A},X} + |\mathcal{B}| \cdot d_{\mathcal{B},X}}{|\mathcal{A}| + |\mathcal{B}|}$$

The UPGMA algorithm produces rooted dendrograms and requires a constant-rate assumption - that is, it assumes an ultrametric tree in which the distances from the root to every branch tip are equal. When the tips are molecular data (i.e., DNA, RNA and protein), the ultrametricity assumption is called the molecular clock.

Create an Object with GEA-NZ

GEA-NZ with the change made in annexe will be use see the taxonomy itself for have the leaf of the taxonomy. To adapted to MISP the GEA-NZ will be transform into MISP J-SON see the documents GEA-json for implementation.

- All new object will be added dynamically to the directory and class object in consequence of applying the methodology explained below

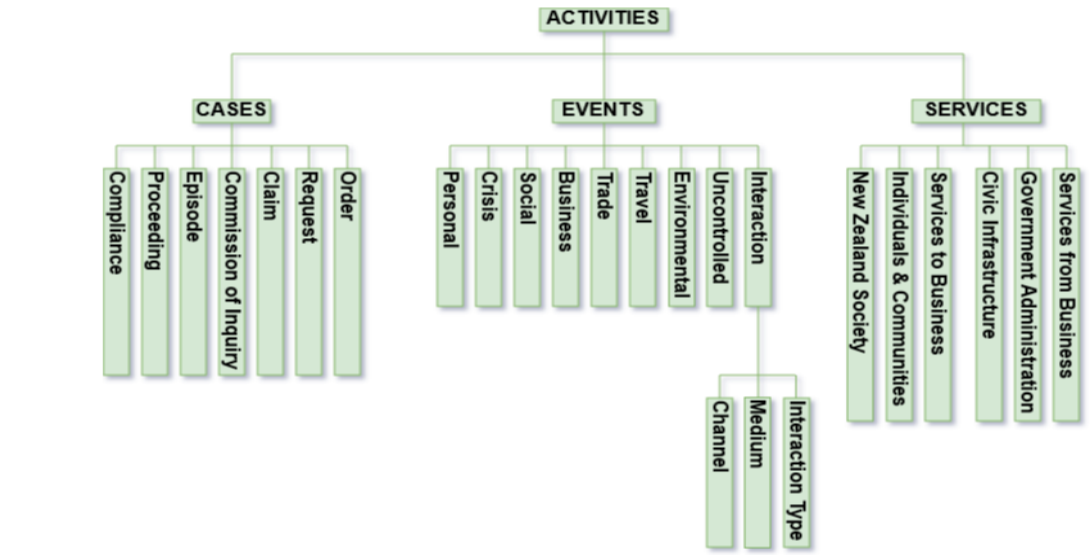
How It Will Works?

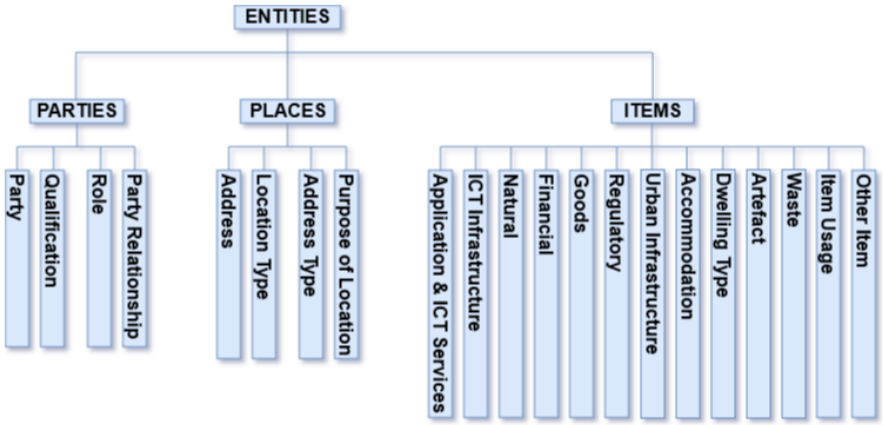
- For each type of object wanted like Job, Activities, Items
    - A table will be made and use as a template to made a Multi-level-model representation which each tables will have independency between columns
        - For structural instances, we can make a table on the the ERCOT Faceted Classification example
        - For type of activities, we can make a table on the "Architecting an Enterprise content management Strategy" example
- Each column of the table template will be fill with the GEA-NZ
- Once each columns are filled, a dissimilatory measure can be made
- Thanks to the dissimilatory measure, we can fill a graphics with points of object instance
- Each points could be regrouping with the K-means method, it will make our object class
- A matrix grouping all the object class with each distance among them is made
- Thanks to the matrix, once created, we can apply the UPGMA method in order to create the clustering dendrogram.
- Now complete directory grouping all the instance object and their class could be made to simplify the creation of further users.

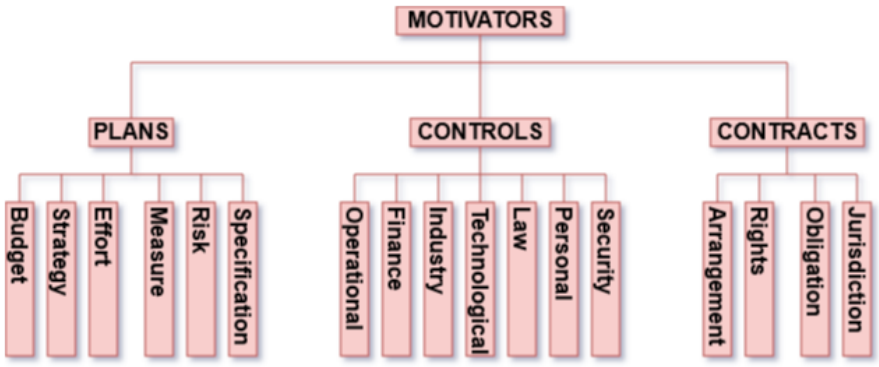
Applied It on MISP ?

- Thanks to the clustering dendrogram and the directory a user could made and choose a job created or modify an existent object.
- To simplify the creation of user, an alternate survey could be made as the ICB Version 2: IPMA Competence Baseline methods
- MISP present many different events, for each representative event, we will must make a template object for each particular type in order to represent it with this methodology.
    - Bugyo and risk governance are added to apply a security approach to the taxonomy

Annexe 2 : GEA-NZ : Government Entprise Architecture New -Zealand





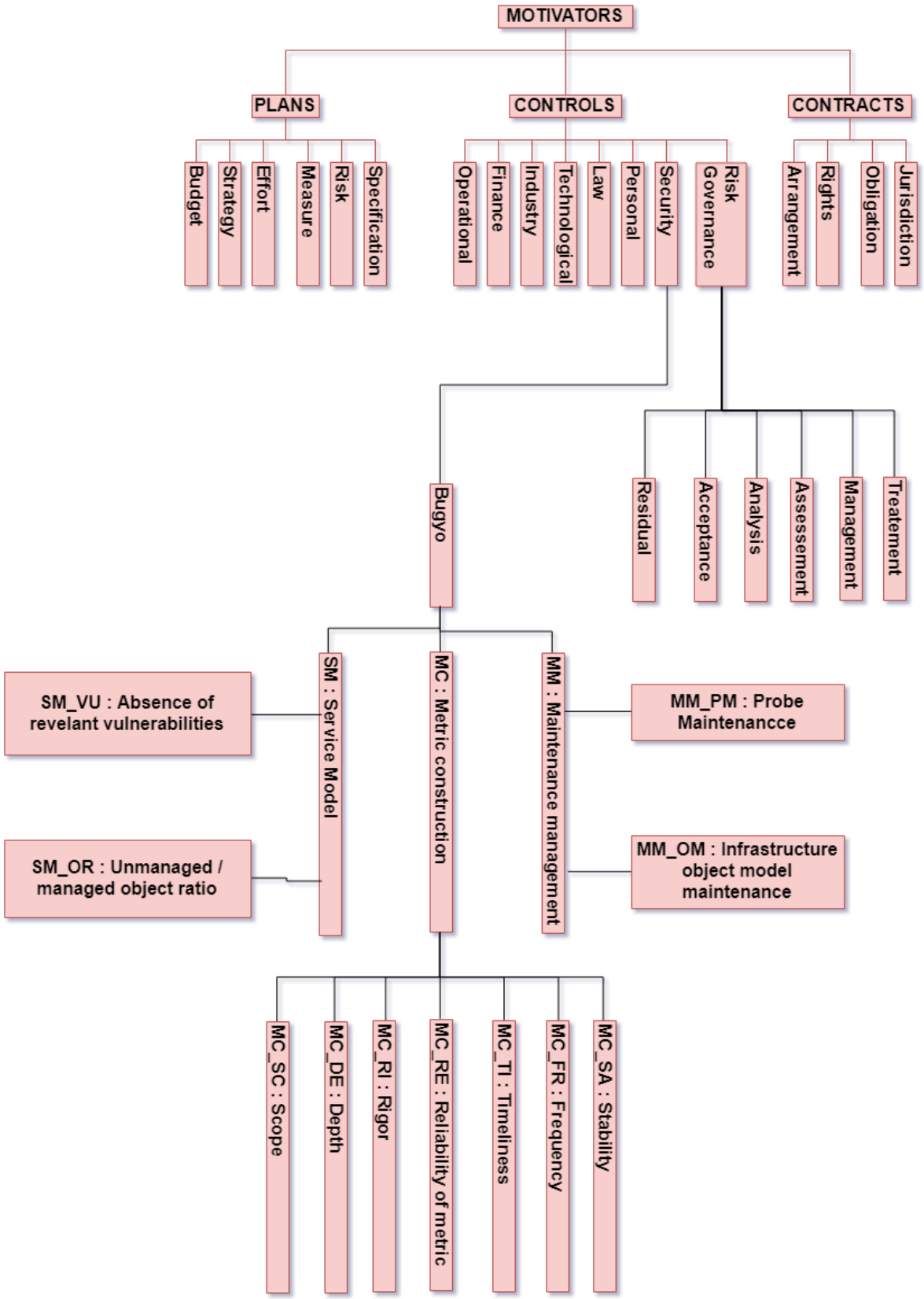


Annexe 2 : Generical Taxinomy JSON


```
{  namespace": "Title of your Taxonomy",
 "description": "Taxonomy to classify the information of ... .",
 "refs": [
   "url ... "
 ],
 "version": 1,
 "predicates": [
   {
     "value": "Name1 ",
     "expanded": "Expanded Name 1 ",
     "description": "TDesciption of the name 1"
   } ,
   {
     "value": "Name2 ",
     "expanded": "Expanded Name 2 ",
     "description": "TDesciption of the name 2"
   }
 ],
 "values": [
   {
     "predicate": "Name1",
     "entry": [
       {
         "value": " value1 ",
         "expanded": "Def Value1"
       } ,
       {
         "value": "value2",
         "expanded": "Def Value2"
       }
     ]
   } ,
   {
     "predicate": "Name2",
     "entry": [
       {
         "value": " value1 ",
         "expanded": "Def Value1"
       } ,
       {
         "value": "value2",
         "expanded": "Def Value2"
       }
     ]
   }
 ]}
```

Annexe 3 : Modification of GEA-NZ

Bugyo and a risk gouvernance have been added



Bibliograph :

[Web] https ://www.misp-project.org.

[Mr Djamel KHADRAOUI,] Mr Djamel KHADRAOUI. cours Bugyo.

[Chris Barden and Amy Lofton and Aubrey Hale and Jenny Ammerman and Claudette Lloyd, 2015] Chris Barden and Amy Lofton and Aubrey Hale and Jenny Ammerman and Claudette Lloyd (2015). Amassing a Practical Enterprise Taxonomy. In ARMA Houston Spring Conference.

[Abhishek Kumar, 2014] Abhishek Kumar (september 2014). Architecting an Enterprise Content Management Strategy : A Four-Pillar Approach.

[A Presentation to Calgary ARMA, 2007] A Presentation to Calgary ARMA (April 2007). Functional Classification Taxonomies.

[The World Bank IBRD IDA, 2016] The World Bank IBRD IDA (2016). Sector Taxonomy and definitions.

[Web] Taxonomy Code Mapping.

[G. Gaupin and H. Knöpfel and P. Morris and E. Motzel and O. Pannenbäcker, 1999] G. Gaupin and H. Knöpfel and P. Morris and E. Motzel and O. Pannenbäcker (1999). ICB.

[The Department of Internal Affairs, 2015] by the Department of Internal Affairs, P. (2015). GEA-NZ v3.1 Business Reference Model and Taxonomy.

[Nezlek, 2015] Nezlek, J. (2015). An Introduction to Multilevel Modeling - basic terms and research examples.

[Web] https ://en.wikipedia.org/wiki/K-means.

[Mr Blansche, 2019] Blansche, A. (2019). M1 AI Course 04 / data search course 05

[Web] https ://en.wikipedia.org/wiki/UPGMA.

[S. Garcia-Vallve, J. Palau and A. Romeu, 1999] S. Garcia-Vallve, J. Palau and A. Romeu (1999). Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in Escherichia coli and Bacillus subtilis. Molecular Biology and Evolution 16(9) :1125-1134.