Knowledge Management in Geospatial Information Context. A Preliminary Statistical Approach - A Case Study

M. FILOMENA TEODORO
CEMAT, Instituto Superior Técnico
Av, Rovisco Pais, 1, 1048-001 Lisboa,
and
CINAV, Naval Research Center
Portuguese Naval Academy
Alfeite, 1910-001 Almada
PORTUGAL
maria.alves.teodoro@marinha.pt

ANACLETO CORREIA
CINAV, Naval Research Center
Portuguese Naval Academy
Alfeite, 1910-001 Almada
PORTUGAL
cortez.correia@marinha.pt

PAULO NUNES
CINAV, Naval Research Center
Portuguese Naval Academy
Alfeite, 1910-001 Almada
PORTUGAL
antunes.nunes@marinha.pt

Abstract: Information is a determinant subject in modern organization operations. The success of joint and combined operations with organizations partners depends on the accurate information and knowledge flow concerning the operations theatre: provision of resources, environment evolution, markets location, where and when an event occurred. As in the past and nowadays we cannot conceive modern operations without maps and geo-spatial information (GI). Information and knowledge management is fundamental to the success of organizational decisions in an uncertainty environment. The georeferenced information management is a process of knowledge management, it begins in the raw data and ends on generating knowledge. GI and intelligence systems allow us to integrate all other forms of intelligence and can be a main platform to process and display geo-spatial-time referenced events. Combining explicit knowledge with peoples know-how to generate a continuous learning cycle that supports real time decisions mitigates the influences of fog of everyday competition and provides the knowledge supremacy. Geo-spatial information and intelligence systems allow us to integrate all other forms of intelligence and act as a main platform to process and display geo-spatial-time referenced events. Combining explicit knowledge with person know-how to generate a continuous learning cycle that supports real time decisions, mitigates the influences of fog of war and provides the knowledge supremacy. These investigation describes the analysis done after the construction and application of a questionnaire and interviews about the GI and intelligence management in a military organization. The study intended to identify the stakeholders requirements for a military spatial data infrastructure as well as the requirements for a future software system development

Key–Words: Geographic Information System, Information and Knowledge Management, Geospatial Information, Geospatial Intelligence, Service Oriented Architecture

1 Introduction

Information is a fundamental resource for military organizations. The success of joint and combined military operations, depends on the existence of intelligence and knowledge about the battlefield in which the forces act: environment, disposal of resources, location of targets, and the events that affect operations and management decisions in real time. Military organizations are aware of the importance of spatial data infrastructures (SDI) to increase data access and sharing, at transnational, national and regional level. For these purposes they need to implement geospatial services, and metadata using an interoperable standards-based services, systems and software. This approach allows shared costs of data collection by reusing same datasets with multiple purposes. Geospatial information (GI) is used to support the operational and administrative activities of interest in organizational context. It has a user-oriented operational purpose and aims to improve the knowledge about a specific process or problem, helping to achieve information superiority and assisting decision-makers in a complex and changing environment. Although with different purposes, intelligence needs information and human tacit knowledge. This work addresses the requirements of interoperability and sharing of GI and geospatial intelligence (GINT) among producers, analysts and consumers. Through the characterization of the current situation it was intended to elicit the different components of the knowledge management cycle. The analysis was done by assessing the current structure and the GINT production capacity. The study also intended, the elicitation of the stakeholders perception of the role and needs from an SDI organization, in order to define the requirements for a software development project. For the case study presented, the objectives were defined, they oriented the questions elaborated in the sense of the perspective analysis of the object of study in order to perceive the current state of development of the phenomenon, identify the needs and define a proposal to improve the capacity of information sharing and geospatial intelligence. The study is presented in six sections. In the next two sections we discuss the details of the questionnaire submitted to the collaborators in the study. The third and fourth sections detail the data issues and the statistics approach. In the last two sections, we present and analyse the results and draw some conclusions.

2 Working Approach to Collect the SDI Stakeholders

The term SDI (Fig. 1) can be define as the relevant base collection of technologies, policies and institutional arrangements that facilitate the availability of and access to spatial data [6]. This common infrastructure requires the combination of multiple databases and software tools linked in a network that makes available to users all spatial data for discovery, reuse and information or knowledge creation by integration at all levels.

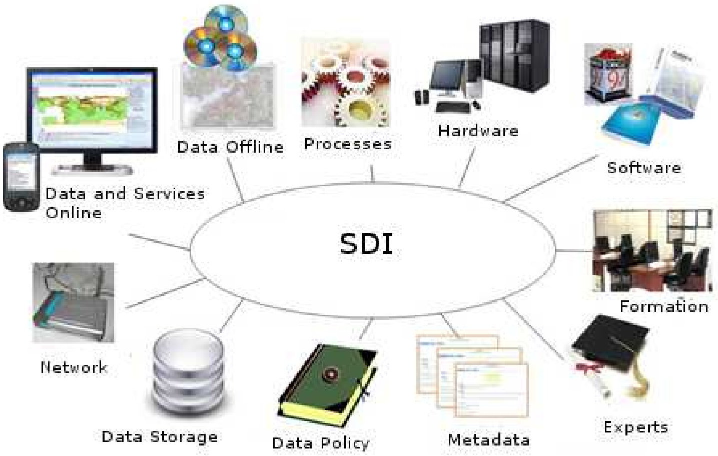


Figure 1: Spatial Data Infrastructure.

An example of this kind of approach could be found in the INSPIRE Directive and the e-navigation initiative. This concept embraces the following elements:

1. Integration processes of technologies, policies, standards, organizations and people;
2. The structure of working practices and relationships across data producers and users;
3. The hardware, software and information technology components necessary to support all processes.

The first step (Fig. 2) in the software development project [12] for the SDI implementation was gathering the users needs and stakeholders requirements, given the inclusive and collaborative nature of the project.

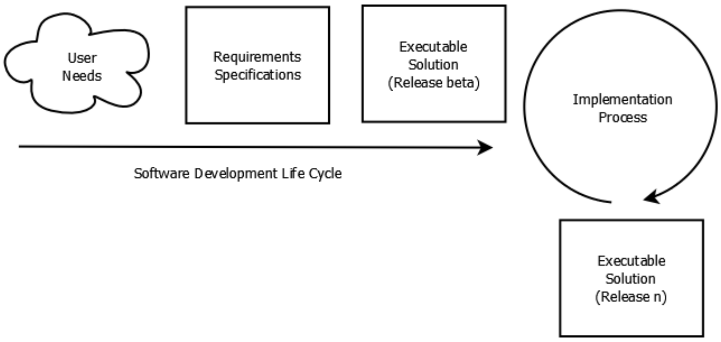

User Needs
Requirements Specifications
Executable Solution (Release beta)
Implementation Process
Software Development Life Cycle
Executable Solution (Release n)

Figure 2: Software Development Life Cycle.

Other approach that did not considered users participation would be condemned to failure. Users needs were surveyed by questionnaires. Other stakeholders (producers) requirements were surveyed by interviews. The collected data was used to define the requirements and a possible solution for SDI and the importance of edify this GI capacity of the military organization.

3 Data

The elicitation of software requirements was based on the survey, interviews and the reference documentation related to SDI development and geospatial data and services interoperability standards.

The survey was applied between January and February 2016, using a web based application (google forms). The sample was drawn according the process described in [5], comprising 103 validated questionnaires, from a population of more than 660 collaborators. There were at least 30 questions per questionnaire, each one with 6 possible ordinal answers.

Qualitative data was also collected from 5 interviews with representatives of main producers of GI, and candidates to provide GI and GINT in the future SDI environment. The elicitation of software requirements was based on the survey, interviews and the reference documentation related to SDI development and geospatial data and services interoperability standards. The collected data was processed through the

Excel Analysis ToolPak plug-in and the ExcelAction Stat plug-in. The statistical approach was completed using SPSS (version 20).

4 Preliminary Statistical Approach Procedure

For elicitation of SDI main requirements, two statistical methods were applied to collected data: the univariate analysis, described in [8, 11], and the exploratory factor analysis described in [9]. The results of FA are still being analysed and extended. Such details will be published in a later article.

The first step in our analysis was to organize the data and get the usual measures of localization, variability and association, using descriptive statistics techniques.

Secondly, the idea was to find some evidence of association between the multiple variables in the study. With this aim, and taking into consideration the ordinal nature of data, we started by the computation of the correlation coefficient Spearmans rho [5, 7]. The non-parametric rank correlation coefficient is computed using (4) where the $r(X_i)$ and $r(Y_i)$ are the ranks of observations X_i and Y_i, respectively, for $i = 1, \ldots, N$, with N the total number of observations:

$$\begin{aligned} \rho = & \frac{\sum_{i=1}^{N}(r(X_i)-\bar{r}(X))(r(Y_i)-\bar{r}(Y))}{\sqrt{\sum_{i=1}^{N}(r(X_i)-\bar{r}(X))^2\sum_{i=1}^{N}(r(Y_i)-\bar{r}(Y))^2}} \quad (1) \\ \bar{r}(X) = & \sum_{i=1}^{N}\frac{r(X_i)}{N}, \\ \bar{r}(Y) = & \sum_{i=1}^{N}\frac{r(Y_i)}{N}. \end{aligned}$$

The expression (4) can be rearranged by algebraic manipulation into (2):

$$\begin{aligned} \rho &= 1-\frac{\sum_{i=1}^{N}(d_i))^2}{N^2(N-1)}, \quad (2) \\ (d_i)^2 &= (r(X_i)-r(Y_i))^2, \quad \text{i=1\ldots,N.} \end{aligned}$$

Based on Spearman rho statistic distribution (chi-square) is possible to calculate the *p-value* associated to the hypothesis of no association versus association. Some times this test does not produce adequate decisions. To mitigate that, the statistic test T (3) with a $t-student$ distribution is often used.

$$T = \frac{\hat{\rho}}{\sqrt{\frac{1-\hat{\rho}^2}{N-2}}}, \quad (3)$$

where $\hat{\rho}$ is the estimator of Spearmann coefficient. A small $p-value$ means a strong association between variables.

After the correlation analysis the authors conduct a non-parametric analysis [5] based on Wilcoxon (W) and Kruskal-Wallis (K-W) tests. The W test is used to compare the median θ of the population with a reference value k,

$$\begin{aligned} H_0 : \theta = &\quad k \\ H_1 : \theta \neq, &\quad < \ or > k \end{aligned}$$

where the distribution of W statistic test presented bellow is asymptotically Gaussian.

$$Z = \frac{\min(S^+,S^-)-(\frac{N(N+1)}{4})}{\sqrt{\frac{N(N+1)(2N+1)}{24}-\sum_{i=1}^{g}\frac{e_i^3-e_i}{48}}} \mathrel{\dot\sim} N(0,1), \min(S^+,S^-) \leq E(s) = \frac{1}{2}\sum_{i=1}^{N} r_i$$

Another test, to compare multiple samples, was used. The KW test is a nonparametric method for testing whether different samples are originate from the same distribution. The median of each subgroup is compared with each other. KW test allows to decide if data has an identically shaped and scaled distribution for all groups (there is no difference in median), then the null hypothesis is that the medians of all groups are equal, either if at least one population median of one group is different from the population median of at least one other group. In KW rank test the hypothesis are validated (or not) using an asymptotically Gaussian distributed test statistic, whose expression is given by (4),

$$KW = \quad (N-1)\frac{\sum_{i=1}^{N}N_i(\bar{r}_{i\cdot}-\bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{N}N_i(r_{ij}-\bar{r})^2}, \quad (4)$$

where

- N_i is the number of observations in group i, $i = 1\ldots,g$;
- g is the number of groups, r_{ij} is the rank of observation j in group i;
- $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{N_i} r_{ij}}{N_i}$, the average rank in group i;
- $\bar{r}$ its the average rank of all observations;
- N is the number of observations in all groups.

Harman [4] published with some detail an extension of Spearmans two factor theory and develop the foundation for the mathematical principals of factor analysis (FA). The main application is to get a reduced number of variables from an initial big set of variables by identification of variables that measure similar things [13]. In the text of [9] the authors presents a description of a preliminary exploratory factor analysis. The results obtained in the study are still under evaluation.

The FA technique is larged used to reduce data. The purpose is to get a reduced number of variables from an initial big set of variables to save time and facilitate easier interpretations [4]. The FA computes indexes with variables that measures similar things. There are two types of factor analysis: exploratory factorial analysis (EFA) and confirmatory factorial analysis (CFA) [2]. It is called EFA when there is no idea about the structure or the dimension of the set of variables. When we test some specific structure or dimension number of certain data set we name this technique the CFA. There are various extraction algorithms such as principal axis factors, principal components analysis or maximum likelihood (see [9] for example). There are numerous criteria to decide about the number of factors and theirs significance. For example, the Kaiser criterion proposes to keep the factors that correspond to eigenvalues greater or equal to one. In the classical model, the original set contains p variables $(X_1, X_2, \ldots, X_p)$ and m factors $(F_1, F_2, \ldots, F_m)$ are obtained. Each observable variable X_j, $j = 1, \ldots, p$ is a linear combination (5) of these factors:

$$X_j = \alpha_{j1}F_1 + \alpha_{j2}F_2 + \cdots + \alpha_{jm}F_m + e_j, \qquad j = 1, \ldots, p, \tag{5}$$

where e_j is the residual. The factor loading α_{jk} provides an idea of the contribution of the variable X_j, $j = 1, \ldots, p$, contributes to the factor F_k, $k = 1, \ldots, m$. The factor loadings represents the measure of association between the variable and the factor [5, 8].

FA uses variances to get the communalities between variables. Mainly, the idea of extraction is remove the largest possible amount of variance in the first factor. The variance in observed variables X_j which contribute to a common factor is defined by communality h_j^2 and is given by equation (6)

$$h_j^2 = {\alpha_{j1}}^2 + {\alpha_{j2}}^2 + \cdots + {\alpha_{jm}}^2, \qquad j = 1, \ldots, p. \tag{6}$$

According with the author of [2], the observable variables with low communalities are often dropped off once the basic idea of FA is to explain the variance by the common factors. The theoretical common factor model assumes that observables depend on the common factors and the unique factors being mandatory to determine the correlation patterns. With such objective the factors/components are successively extracted until a large quantity of variance is explained. After the extraction technique be applied, it is needed to proceed with the rotation of factors/components maximizing the number of high loadings on each observable variable and minimizing the number of factors. In this way, there is a bigger probability of an easier interpretation of factors 'meaning'.

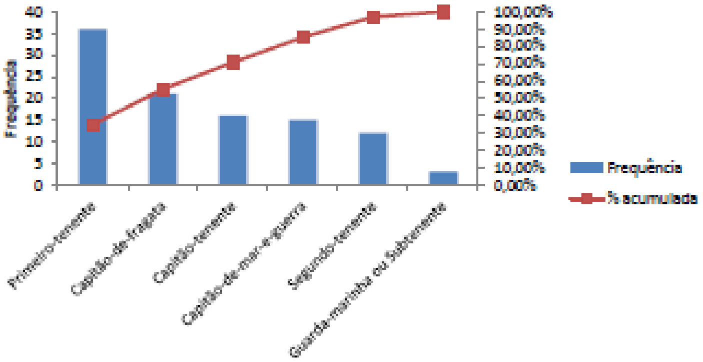


Figure 3: Number of responses per rank.

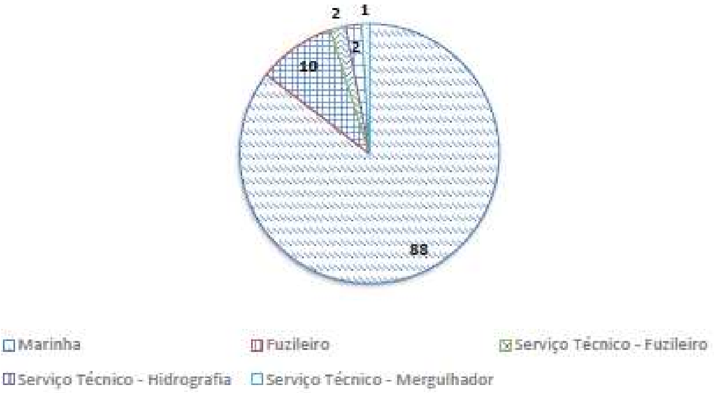


Figure 4: Number of responses per class.

5 Results from quantitative and qualitative data

By first, all data was organized and classified using the descriptive statistics. The summary of data can be found in Figs. 3 and 4. In Fig. 3 are displayed the number of responses per profissional rank. In Fig. 4 are displayed the number of responses per profissional classe.

Comparing all the answers using the Spearman rho we obtained a matrix with Spearman rho values and the significant associations. The results show a strong correlation between the actual users perception

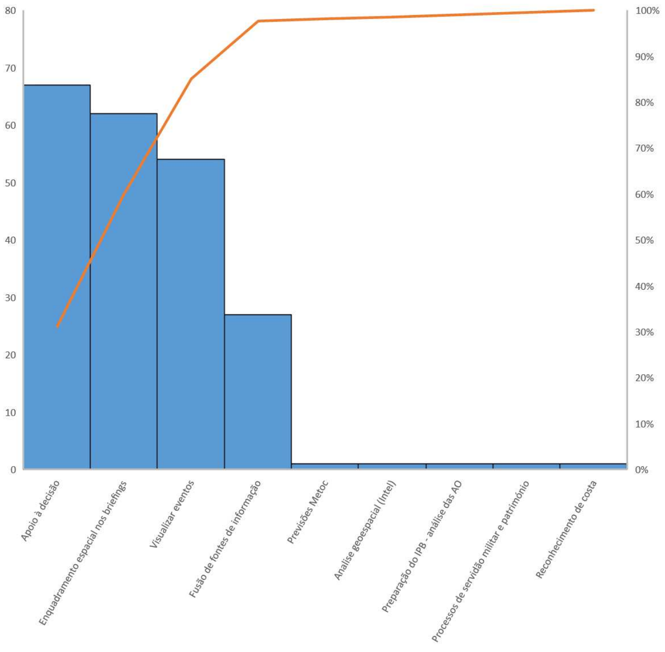


Figure 5: Types of GEOINF used.

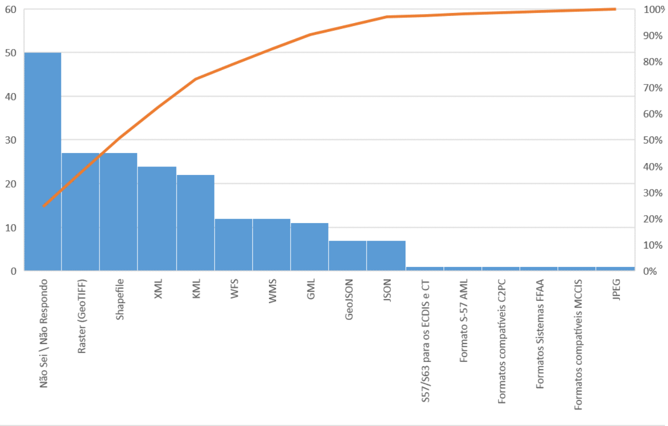


Figure 6: Types of used formats.

of importance of GI and GINT and the importance tend of this resource for the future.

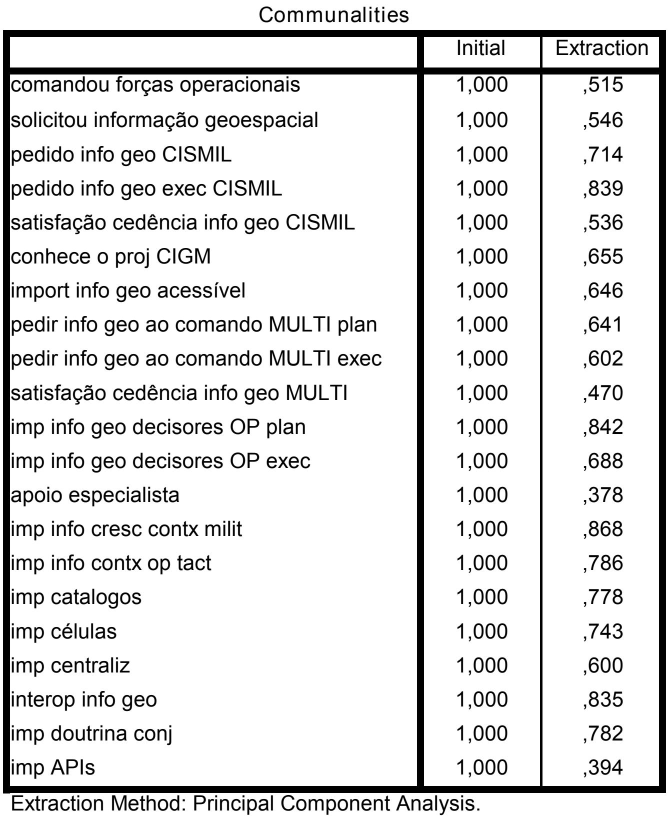
Data gathered from survey allowed to confirm the importance of GI and GINT for end users and decision makers. The statistics significance were tested using the non-parametric tests of Wilcoxon and Kruskal-Wallis [3, 4]. In KW test, the sub-data sets are aggregate by the professional experience of users; in the present case by their rank in organization. It was statistically validated that users indeed acknowledge the importance of GI and GINT and understand the influence of this resource in the military decision process.

The survey allowed also to collect information regarding the datasets that users mostly require: reference and cartographic data, meteorological and oceanographic dynamic data and open crowdsourced data. A summary about the types of used GEOINF can be found in Fig. 5. Regarding interoperability standards, users opted to the Open Geospatial Consortium data and services formats as can be found in Fig. 6.

Furthermore, according with [1, 5, 10] an exploratory factor analysis, presented in [9], allowed the identification of the factors that adequately explain the variance of users requirements. The selection of such factors used several criteria, just like in [3]. Summarizing the results of the exploratory FA approach, the FA was applied to the model allowing the extraction between three and six factors, which together are able to explain from 46% and 64.5% of the total variance of the model. To verify the consistency of the original data the Kaiser-Meyer-Olkin test (KMO), with an index equal to 0.77, standing at a good interval, which evidenciates adequacy of factorial analysis apprroach. In performing a second test, the Bartlett's Test of Sphericity (BTS), it has been found that it is unlikely that the correlation matrix is a identity. This is represented by a high index generated by the BTS test (1070) and a zero $p-value$. Table 1 shows the commonalities for each variable. It can be noticed that only 3 of the 21 variables used had the communalities values below 0.500, that is, in 21 variables, more than half of the variance of each variable is reproduced by the common factors and for 7 variables this value is above 0.750. It is possible to verify the factor loads and identify the coefficients of the columns that represent the relationship between each of the variables and their respective factors. If we takinto consideration e the factor loads with the highest value for the variables, we have a higher chance to give a ”meaning” to some factor..A detailed description of such FA approach will be done later.

The interviews treatment allowed the understanding of current situation regarding military capability elements, namely, about the network that consumers, producers and the coordinator use to transfer geospatial information (off-line analog and digital data). The group of experts were in fact aware of the importance of geospatial information sharing. They also identified the overall elements of the military capability [14] in the network (Fig. 7), and recognized the importance of standards adoption, such the ones in used in multinational organizations, such as NATO and European Union, in order to increase the interoperability and reduce the costs of software development projects.

So, the conducted interviews allowed a comprehensive knowledge about the actual situation and how, in the future, the solution should be designed to in-



Communalities

	Initial	Extraction
comandou forças operacionais	1,000	,515
solicitou informação geoespacial	1,000	,546
pedido info geo CISMIL	1,000	,714
pedido info geo exec CISMIL	1,000	,839
satisfação cedência info geo CISMIL	1,000	,536
conhece o proj CIGM	1,000	,655
import info geo acessível	1,000	,646
pedir info geo ao comando MULTI plan	1,000	,641
pedir info geo ao comando MULTI exec	1,000	,602
satisfação cedência info geo MULTI	1,000	,470
imp info geo decisores OP plan	1,000	,842
imp info geo decisores OP exec	1,000	,688
apoio especialista	1,000	,378
imp info cresc contx milit	1,000	,868
imp info contx op tact	1,000	,786
imp catalogos	1,000	,778
imp células	1,000	,743
imp centraliz	1,000	,600
interop info geo	1,000	,835
imp doutrina conj	1,000	,782
imp APIs	1,000	,394

Extraction Method: Principal Component Analysis.

Table 1: Communalities.

crease the processes of data, information and knowledge sharing.

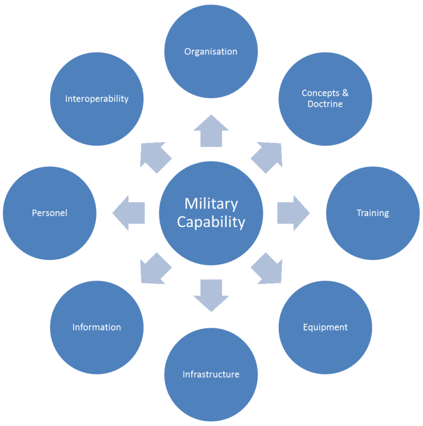


Figure 7: Military Capability Elements.

6 Conclusions and Final Remarks

In this work users survey data analysis, as well as treatment of experts interviews were used to extract the users requirements for software development projects. The approach allowed the elicitation of current state of geospatial information and geospatial intelligence sharing in an organization ecosystem.

In the elicitation process decision makers recognize the importance of geospatial information to increase speed and quality of decision process. This comes aligned with acknowledge that, in the current days, technology speeds up the sharing processes and organizations needs to find new ways to achieve horizontal and vertical integration by information sharing and interoperability between systems. Since interoperability depends on standardization agreements, the best way to achieve is the adoption of metadata, standard data and services from Open Geospatial Consortium, International Standardiza-tion Organization, NATO, European Union INSPI-RE, and International Hydrographic Organization.

In conclusion, both users and experts acknowledge on the adoption of a SDI conceptual model (Fig.8) with a network of distributed military capability elements available to producers, coordinators as well as consumers of geospatial information.

Acknowledgements: This work was supported by Portuguese funds through the *Center of Naval Research* (CINAV), Naval Academy, Portuguese Navy, Portugal and the *Center for Computational and Stochastic Mathematics* (CEMAT), *The Portuguese Foundation for Science and Technology* (FCT), University of Lisbon, Portugal, project UID/Multi/04621/2013.

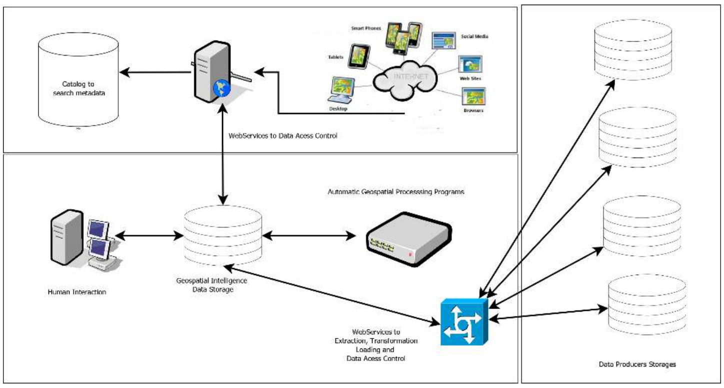

Catalog to search metadata
Smart Phones
Social Media
Tablets
INTERNET
Web Sites
Desktop
Browsers
WebServices to Data Acess Control
Automatic Geospatial Processing Programs
Human Interaction
Geospatial Intelligence Data Storage
WebServices to Extraction, Transformation Loading and Data Acess Control
Data Producers Storages

Figure 8: SDI Conceptual Model.

References:

[1] T.W. Anderson, *An Introduction to Multivariate Analysis*, Jonh Wiley & Sons, New York, 2003.

[2] D. Child, *The Essentials of Factor Analysis*, Continuum International Pub. Group, New York, 2006.

[3] J.F. Hair, R.E. Anderson, R.L. Tatham, W.C. BLACK, *Multivariate Data Analysis*, 4^{th} ed., Prentice Hall, New Jersey, 1998.

[4] H.H. Harman, *Modern Factor Analysis*, Univ. of Chicago Press, Chicago, IL., 1976.

[5] J. Marco, *Análise Estatística com o SPSS Statistics*, ReportNumber, Pêro Pinheiro, 2014.

[6] New Zeland Spatial Office, *Spatial Data Infrastructure Cookbook*, Global Spatial Data Infrastructure Association, Melbourne, 2012. [cited 2016 november] Available from: `http://gsdiassociation.org/images/publications/cookbooks/SDI_Cookbook_from_Wiki_2012_update.pdf`.

[7] A. Mood, *Introduction to the theory of Statistics*, McGraw-Hill Inc., Auckland, 1984.

[8] P.A. Nunes, *Gestão da informacão e conhecimento - desafios para a marinha: processos de recolha, análise, gestão e transferência de inteligência geoespacial*, Monography, Instituto Universitário Militar, 2016.

[9] P.A. Nunes, A. Correia, and M.F. Teodoro, *Information Gathering, Management and Transfering for Geospatial Inteligence*, in ICNAAM 2016, AIP Conference Proceedings, edited by T. Simos et al., American Institute of Physics, Melville, NY, (in press).

[10] S. Sharma, *Apllied multivariate techniques*, John Wiley & Sons, New York, 1996.

[11] A.C. Tamhane and D.D. Dunlop, *Statistics and Data Analysis: from Elementary to Intermediate*, Prentice Hall, New Jersey, 2000.

[12] F. Tsui, *Managing Software Projects,* Jones and Bartlett Publishers, Inc., Burlington, MA, 2004.

[13] A.G. Young and S. Pearce, A Beginners Guide to Factor Analysis: Focusing on Exploratory Factor Analysis, *Tutorials in Quantitative Methods in Psychology* 9(2), 2013, pp. 79-94.

[14] Y.Yue and M. Henshaw, An holistic view of UK military capability development, *Defense and Security Analysis* 25(1), 2009, pp. 53–67.