JHI JOURNAL OF HEALTH INFORMATICS

Health Information System (HIS) security standards and guidelines history and content analysis

Estudo cronológico e de conteúdo de normas de segurança da informação para Sistemas de Informação em Saúde (SIS)

Estudio cronológico y del contenido de normas de seguridad de información para los Sistemas de Información Sanitaria (SIS)

Marcelo Antonio de Carvalho Junior[1], Cristina Lúcia Feijó Ortolani[2], Ivan Torres Pisa[2]

Keywords: Information Systems; Standards; Computer Security

ABSTRACT

Objective: This article provides for identification and content study of main standards and guidelines used to support Health Information System (HIS) development. **Method:** Standards deemed used as reference by SIS developers were list. The different cited standardization organizations' production was assess for history and content analysis. Cited documents were acquire and its contents automatically extracted for study. We manually listed all references to outer content declared within assessed documents. Then, we apply different text analysis methods to decompose, link and correlate the content to disclose inner relationships. **Results:** Document similarity analysis on standards resulted between 5% to 89%. A total of 440 outer-connections were found. The most influential documents according to Betweeness-Centrality and average-path from these connections were casted. The density found on this graph is 0,6%. **Conclusion**: This study provided for a better understanding of existing HIS standards.

Descritores: Sistemas de Informação; Normas; Segurança Computacional

RESUMO

Objetivo: O objetivo deste trabalho é identificar os principais documentos utilizados como referência para construção de Sistemas de Informação em Saúde (SIS) e estudar seu conteúdo. **Método:** Identificamos referências utilizadas por desenvolvedores de SIS. Listamos e identificamos cronologicamente a produção das entidades normativas citadas. Adquirimos os documentos citados e em seguida extraímos automaticamente seu conteúdo para estudo. De forma manual, listamos todas as referências internas desses documentos a outras normas. Aplicamos então diferentes métodos de análise de texto para resumir, decompor e correlacionar o teor do conjunto. **Resultados:** As análises identificaram similaridades entre os documentos, variando de 5% à 89%. Por meio da análise de referências externas, localizamos 440 ligações. As normas mais influenciadoras no conjunto foram elencadas segundo índice Betweeness-Centrality. A densidade dessas ligações entre os documentos é de 0,6%. **Conclusão:** Por meio de estudo histórico e de conteúdo, promovemos um melhor entendimento de normas existentes.

Descriptores: Sistemas de Información; Normalización; Seguridad Computacional

RESUMEN

Objetivo: Este artículo describe el estudio de identificación y contenido de las principales normas y directrices utilizadas para apoyar a lo desarrollo de Sistemas de Información de Salud (SIS). **Método:** Identificamos las normas utilizadas como referencia por desarrolladores de SIS. Enumerados y identificamos por orden cronológico la producción de los organismos reguladores mencionados. Adquirimos los documentos citados y extraemos automáticamente su contenido para estudiar. Manualmente, listamos todas las referencias internas de estos documentos mencionando otras publicaciones. Em seguida, aplicamos diferentes métodos de análisis de texto para resumir, descomponer y correlacionar todo el contenido. **Resultes:** Análisis de similitud documento de estándares resultó entre 5 % a 89 %. Se encontró un total de 440 conexiones externas a ellos. Se descubrió los documentos más influyentes según métrica Betweenes-Centrality. La densidad medida en esta red es del 0,6%. **Conclusión:** Através del estudio histórico y de contenido de normas existentes, logramos a promover su mejor comprensión.

[1] *Pós-graduando em Gestão e Informática em Saúde, Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil.*
[2] *Departamento de Informática em Saúde, Escola Paulista de Medicina, Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil.*

Autor Correspondente: **Marcelo Antonio de Carvalho Junior**
e-mail: **carvalho.junior@unifesp.br**

Artigo recebido: **17/04/2015**
Aprovado: **23/09/2015**

INTRODUCTION AND BACKGROUND

Certifying bodies serve as trusted anchors for Health Information Systems (HIS) quality assertion. Standards and guidelines published by national and international bodies are widely used as common ground for HIS development and are used for system certification. Along with local legislations, country-specific standards and business requirements they are references used for system construction worldwide. They were produced over the last two decades and are updated constantly to cope with emerging technologies, reflect accepted best practices and provide compliance to HIS specific goals. National, regional and international fronts like the American National Standards Institute (ANSI), Institute of Electrical and Electronic Engineers (IEEE), National Electrical Manufacturers Association (NEMA), European Committee for Standardization (CEN), Health Level Seven International (HL7) and International Organization for Standardization (ISO), are largely cited as sources from this content.

The plural production of this content resulted in a diverse variety of documents with different titles and scopes. Even documents coming from the same source can be divided in parts with particular focus making difficult to find and select the appropriate requirement text to follow. Given this multitude, unless strictly dictated by a legislation, a HIS developer is challenged with the decision of what source and document(s) a system should be based on.

Different previous papers explored standards' content from distinct aspects. Buckley[1] provided a view of the IEEE's standards production process. Oksala et al.[2] gave a high level overview of IT standardization. Leistner[3] discussed ANSI's members and experts composition. Johnson et al.[4] discusses the "neutrality" of standardization bodies' members and the impact on standard production. Timothy's paper[5] published last year provides a broad overview and discussion of the immediate history of some of the standards produced by these sources from a US perspective.

This article describes and discusses this content and provides a historical overview of health informatics' focus standards production from ASTM and ISO specific working groups over the years. Using mostly automated methods and unsupervised means, content analysis from standards texts is performed selecting the information security portion from these documents as we investigate:

a) Content superposition degree among standards;
b) Standards documents inter-relationships;
c) Comparisons from ANSI and ISO approaches and main topics discovery.

The following sections are divided as: (a) Methods description, where standard´s text acquisition, extraction (TE) and storage for processing are described as well as the text mining algorithms used to explore the three lines of investigation. (b) Standards and Guidelines chronological perspective, that describes ISO and ANSI' health informatics working groups, discuss production flow and place their publishing over a timeline for analysis; (c) Standards and Guidelines content, discussing main topics found, correlations, text superposition and an outer content analysis based on references to external documents found. (d) At the Discussion section the overall perception of findings and the method execution is discussed. (e) Conclusions.

METHODS

Based on auditees declarations of standards used as base for system production during 16 audits performed by Brazilian Society of Health Informatics (SBIS - http://www.sbis.org.br/), we have selected 36 documents from ASTM and ISO full list for study. Following these criteria, we also added the later versions from Certification Commission for Health Information Technology (CCHIT - http://www.cchit.org/) and SBIS ambulatory requirements totaling 38 documents at our data corpus. These information were collected anonymously (no respondent or HIS information retrieved) after consent using pre-audits declarations and questioner responses from auditees prior to HIS audit. The study approval was obtain from ethical committee under Certificate of Presentation for Ethical Consideration (CAAE) #11933213.5.0000.5505. For standards timeline study, we have downloaded a list of published ASTM and ISO documents containing those cited during audits and all others from health informatics committees at their original websites (http://www.astm.org/ and http://www.iso.org/ respectively). Then, we manually checked for previous versions and updates history to populate a table.

As Hassan and Baumgartner[6], we have then extracted the text content from our corpus (pdf documents set) using open source "Pdf-extract" library (http://labs.crossref.org/). The process used headings and text fields to extract inner requirements titles and requirement texts for MySQL database insertion allowing later content and correlation analysis. Vector Space Model (VSM) and TF-IDF[7-8] algorithms are used during text processing to represent the weight of features (terms) found on our corpus allowing correlation and topic discovery here discussed. Automatic key-phrase extraction techniques is applied using open-source (GNU General Public License) KEA algorithm and maui-indexer implementation in order to visually represent the corpus content into main topics.

For external content interconnection, we manually assessed all documents searching for references within text producing a list for each standard. Ranking values for node (standards) connection to external documents based on Betweeness-centrality[9] metrics are used to locate most influential standards.

STANDARDS AND GUIDELINES HISTORY

Different groups and experts committees around the world are responsible for standard production over the years. Health informatics represent a significant share of standards industry and focus. ISO TC 215 for instance has a huge active committee that is recognized by World

Trade Organization (WTO) for product development. The standards production is related to HIS' structure and architecture, vocabulary and content, storage, security, confidentiality, functionality, messaging and interconnection. Published content is revised/reviewed periodically producing reapproved, reviewed or updated texts. A brief description from this entity and the others assessed in this paper can be seen below:

A. ISO

Created in 1998, the technical committee 215 (TC 215) is the ISO division responsible for health informatics. International, it´s formed by 33 participant countries and internally divided in 10 working groups/joints. Several liaisons were established with 31 different committees and the following organizations: CDISC, COCIR, DICOM, EFPIA, GS1, HON, ICN, IHE, IHTSDO, IMIA, INLAC, ITU, UNECE, WHO, WONCA, mHealth Alliance.

B. ANSI – ASTM

ASTM is an American standardization accredited by ANSI.

Created in 1970, the E31 produced its first standard delivery in 1995. It currently has over 30 approved standards produced by 300 members divided in 6 technical subcommittees. Their production is grouped for annual publishing in the format of Annual Book of ASTM Standards.

ASTM Standards are reviewed in a 5 years window resulting in updated or reapproved versions. Standards that are not revisited for this process within 8 years are automatically withdraw.

C. CCHIT

CCHIT is a non-profit organization established in 2004 based on volunteer efforts of commissions and work groups coordinated by permanent small staff. CCHIT has also been recognized and accredited by National Institute of Standards and Technology (NIST). First versions of standards (certification criteria) was publish in 2008-2009. Ambulatory, Inpatient and Emergency Department are the tree main CCHIT programs followed by supplementary material set which include Cardiovascular Medicine, Child Health, Dermatology, Clinical Research, Oncology and Women's Health. CCHIT has ended its operation last November and no new requirements or updates will be produced.

D. SBIS

Accredited by Medicine Federal Council (CFM), SBIS is a health informatics Brazilian association created in 1986. Its first contribution to a standardization product dates 1999 in partnership with SUS informatics department (DataSUS). The first of the three published HIS' certification standards of its own was made available in 2004. The requirements are produced by a group of experts and a public consultation is performed prior publication.

The standards from those sources totals 124 documents. SBIS and CCHIT publishing (2 documents) are country-specific though and heavily based on local needs and legislations (only ambulatorial requirements were considered for this study). Health Information Portability and Accountability Act (HIPAA) goals are deeply related to CCHIT's requirements[10]. SBIS document is divided in three parts, focus on structure, security and digital signature. As certification bodies, they not only produce standards but also seal approved HIS. Internationally adopted, the document content types from ASTM and ISO can be seen at Table 1.

Table 1 - Published standards and guidelines per sources

Type	Current HIS standards and guidelines	
	ASTM – E 31	ISO – TC 215
Messaging	0	6
Structure	6	38
Security	13	21
Vocabulary	1	3
General	17	17

Considering ASTM and ISO history, we can see at Figure 1 that "waves" of standards publishing comes every 1-2 years and no interaction gap since starting point. The publishing of new documents started five years earlier on ASTM production in comparison to ISO but ceased at 2006 followed only by updates. ISO instead, is producing new content until the present date, also followed by a number of updates. Comparisons aren't straight forward though as in ISO chart there´s no "reapproved" status, those hence also labeled publishing.

Another phenomena that distinguish this graphics is that new names are frequently attributed to updated ISO standards. Thus, the most comparable scenario would be the sum of published plus updated on ASTM chart. By assessing the security content over this two production timeline, is possible to see that ASTM incorporated information security contents since the first year of production, while the first event from ISO occurs only at 2004. This characteristic denotes security as a second thought as the first focus was functionality and architecture. A full timeline list constructed based on this two organization's production including standards titles can be seen depicted at http://telemedicina2.unifesp.br/valorarequisitos/arquivos/fulllist_horizontal_HIS_history.html

The standards production dates though starts way back since there is a formal proposition, draft and revision process that needs to be undertaken before approval. There are specific rules and bureaucratic flow for standards development, including the procedures required to bring a document to publication. In ISO for instance, this process takes 4-5 years average as the formulated standards texts are refined and finally approved to be made available to end users. As result, the standards used to guide HIS does not necessarily contain cutting-edge technology available but the ones widely accepted and consensually adopted. To diminish the potential gap of using "outdated" references, ISO changed the standards publishing workflow since last committee meeting at

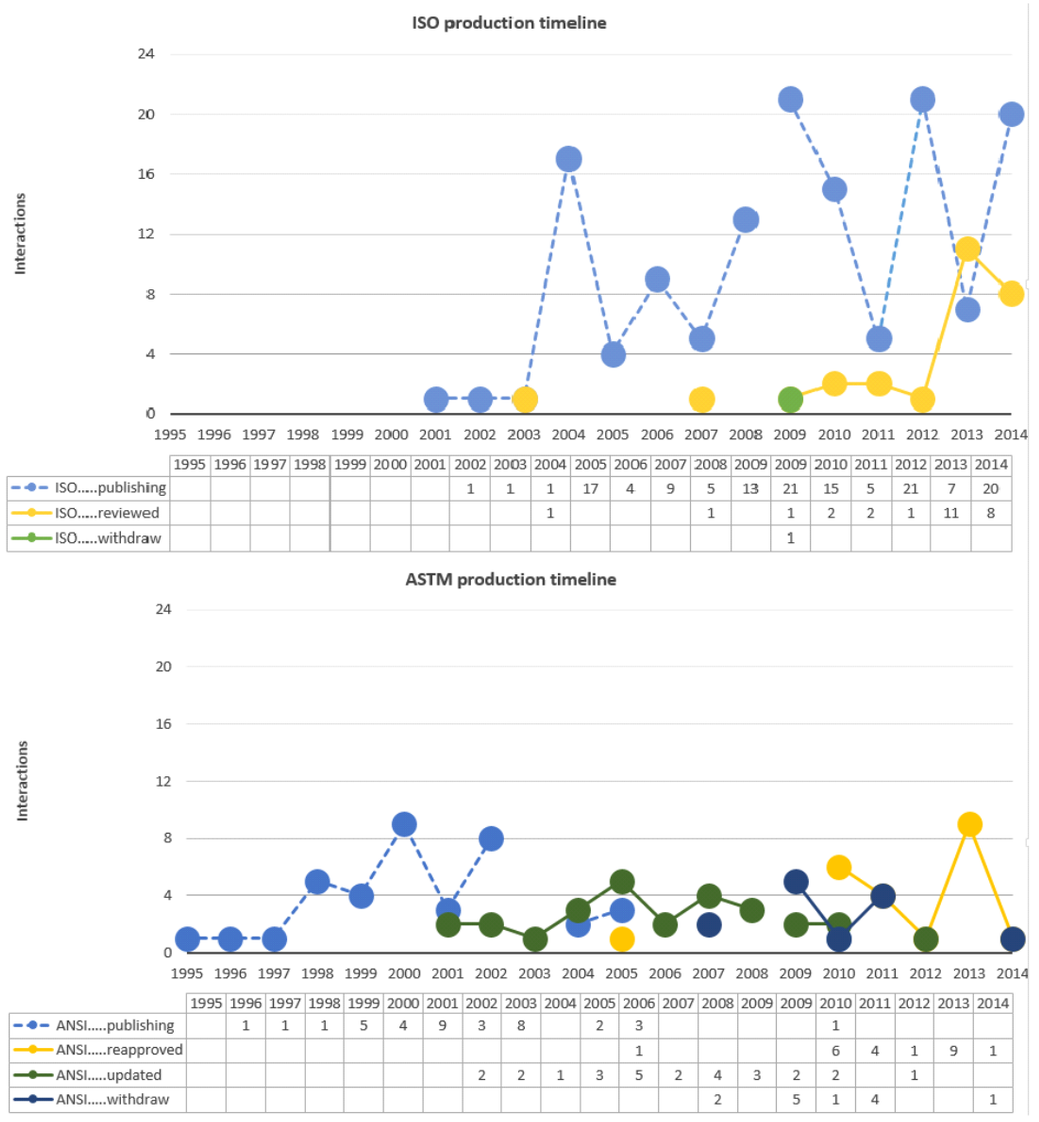

Figure 1 - ASTM and ISO production timeline

Karuizawa – Japan. The normal procedure stages that were composed of 5 different obligatory phases are now shortened to 2 as seen in Figure 2. Optional phase is only required if negative votes are registered at draft voting. The time saving planned by using the reduced framework is 1 year. The number of voting sessions was keep.

STANDARDS AND GUIDELINES CONTENTS

Focused on HIS' security features, a portion of ASTM and ISO collection cited by auditees was acquired for content analysis. The rationale decision for choosing the

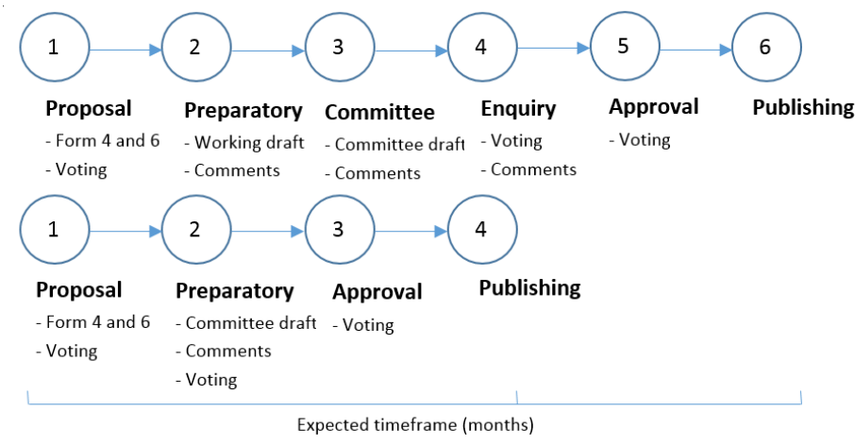

Figure 2 - Reduced ISO framework for standards approval and publishing

requirements selection (studied corpus) shown on Table 2 refers to our experience as auditors. During SBIS audits sessions, the auditee declarations were taken in consideration. Thus, the below table represents the most common texts deemed as reference used for software construction or adaptions for compliance purposes by companies subjected to SBIS certification that shared that information. Also, we believe that all existent HIS requirements available has a certain degree of superposition as they dictate about same topic Figure 3.

Main topics

Key-phrases are important phrases found within a document. Extracting main topics (key-phrases) in the security-related standard collection can help a system developer or other security practitioner to locate the relevant documents that should be used. As in the other studies[10-11], KEA algorithm was used for topic extraction. The resulting table may aid reader in this information lookup task.

We also added a coefficient value to the table that positions the standards in a similarity scale compared to the overall text on our database. The higher the number (obtained by Jaccard similarity function), more common textual terms are used on that particular document. That means that following a low rated standard is likely that one can find and implement the functionalities that are more specific to a system. As seen on below table, many documents refers to common topics.

Similarity comparisons took place considering word-matching only. There are similarities even comparing the two different standardization organizations. Figure 3 shows similarity above 0.25 when comparing pairs of documents against each other. This comparisons resulted between 0.05 and 0.89. As shown, ISO 27799 and ISO 27789 boldly resembles each other although represented by different main topics. Low mapping between different documents usually indicates underived content and therefore should be preferred in terms of broader system feature coverage. This pair-wise visualization is useful for

Table 2 - Studied security-related standards

Standard	Analysis	
	Key-phrase extraction	Jaccard index
SBIS NGS1	access control, EHR components, communication, data access	0.037829216424014
SBIS NGS2	digital certificate, PKI Brazil, authentication	0.017335579016555
SBIS ECF	EHR information, health professional, user identification, legal	0.019423868312757
CCHIT	patient record, authentication, system, services	0.062952898550725
ASTM E2595	certificates, credential, privilege management	0.15802692798617
ASTM E2538	operations, EHR specification, life cycle	0.097541894840861
ASTM E2473	Electronic Health Record, environment, data elements	0.038580310454152
ASTM E2436	data structure, data model, identification	0.034462881376868
ASTM E2369	clinical practitioners , patient care, personal health information	0.041338987935933
ASTM E2212	health information, certificates, CA	0.10100053526578
ASTM E2171	measurement, data quality, scale unit	0.13093768008562
ASTM E2147	audit log, disclosure, audit functions	0.035945155844691
ASTM E2145	business processes, structures, information modelling	0.074607814880389
ASTM E2017	health information, data entry	0.027257380491621
ASTM E1986	data elements, privileges, data Access	0.056573475521884
ISO/TS 25237	pseudonymized data, protection, privacy, identification	0.10565323012311
ISO/TS 21298	roles, privilege management, access management	0.05159138633837
ISO/TR 20514	EHR system, integration, shareable	0.08564252480751
ISO 18308	management, communication, Electronic Health Record	0.067896405484416
ISO 17090-2	CA certificates, certificate type, identity certificate	0.052085477827644
ISO 17090-1	digital certificates, healthcare providers, signed information	0.072507926050974
ISO/TS 22600-1	access control, privilege management, information exchange	0.039774364886565
ISO/TS 22600-2	privilege management, role, authorization	0.041709556552888
ISO 13606-5	Audit Log, communication, EHR data, EHR system	0.022192942726562
ISO/IEC TR 15026-1	assurance , activities, evidence	0.17989047638654
ISO/TS 22600-3	privilege management, access control, security, services	0.047762177296496
ISO/TS 13606-4	EHR system, EHR data, EHR communication, access, audit log	0.05187960637378
ISO/IEC 12207	system requirements, software product, life cycle	0.14361592621567
ISO/TS 21547	EHR archive, security, EHR system	0.081401572857908
ISO/TR 21548	eArchiving, security requirements, metadata, policy	0.073207888994112
ISO/TS 21091	management, object class, directory	0.045291719850126
ISO 27789	audit trails, personal health information, audit data	0.16992629801952
ISO/TS 14441	privacy and security, clinical, audit, access control	0.10836728814257
ASTM E1985	organizational policy, authorization mechanisms, assessment	0.036602437417655
ASTM E1869	privacy, identifiable, health information systems, responsible	0.06085560176226
ASTM E1762	electronic signature, user authentication, identity, signer, attributes	0.080660435623996
ASTM E1384	structure, Electronic Health Record, data elements	0.1264462469634
ISO 27799	information security, risk, information systems, governance	0.12599332976489

implementers that wishes to find the text intersection degree. No significate changes were noticed when applying Wordnet into the comparison process. According to Abrahan and Idicula[12], the use of this synonymous thesaurus for additional similarity computation is restricted to nouns or verbs for mapping. We also consider that some of this effect is due the extensive use of acronyms within technical texts that are not present on Wordnet´s dictionary or even an indicator that the documents are quite consistent in terms of writing style/language. A full list of similarity table can be found at http://telemedicina2.unifesp.br/valorarequisitos/arquivos/Estudo%20similaridade%20requisitos.xlsx .

Although not covered from this research, we believe that not only connections among different standards and similarity content analysis are important but also the other way round. Conflicts within standards were also reported from some auditees as a difficulty for correct security implementation. Additional study is needed to cover that matter.

Outer content relationship

A total of 440 external references were found within studied corpus. That is an indicator that complimentary reading may be needed to fulfill a certain feature understanding and hence correct implementation. Some of the references cited within requirement texts or at the reference topic produces a cascading effect and in some cases a loop. For instance, the NIST SP 800-53 reference is cited many times in CCHIT requirements but when the full reference is assessed is quite common to find that it belongs somewhere else, like the Open System

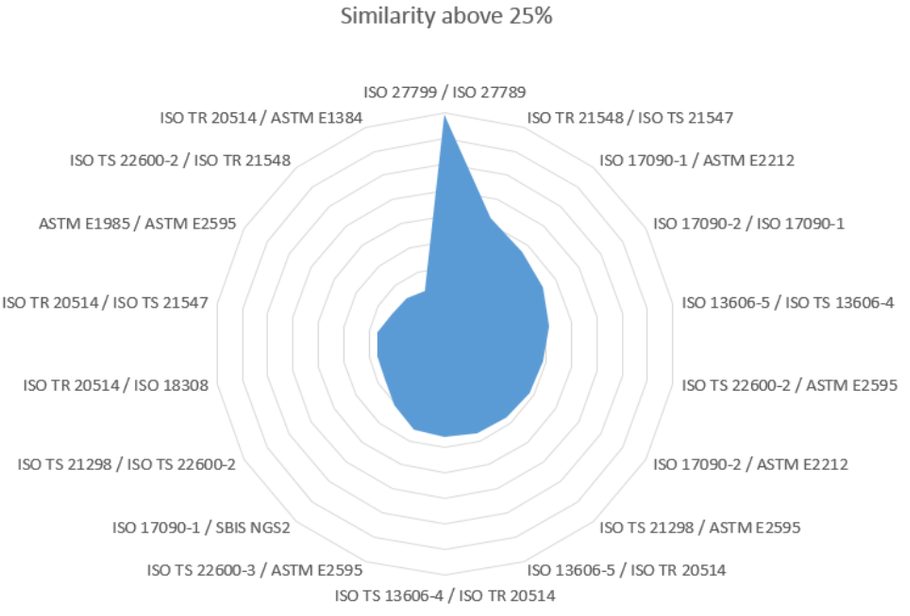

Figure 3 - Over 25% similarity between studied standards

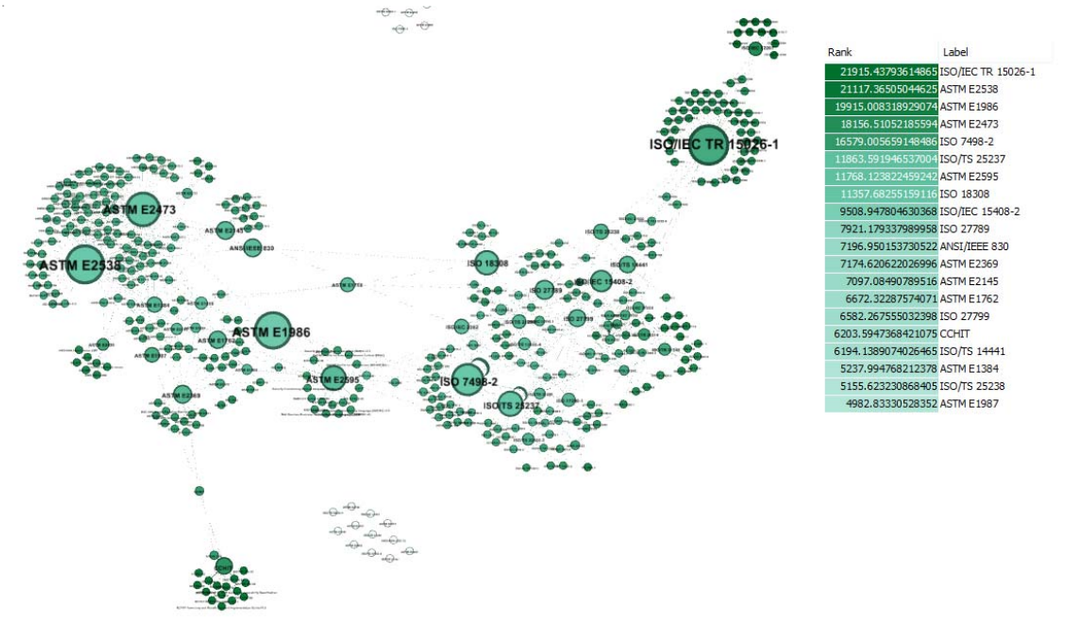

Figure 4 – Top 20 reference Betweeness Centrality interconnection

Architecture opensecurityarchitecture.org and others.

For the sake of relationship establishment and graphical representation, we´ve omitted the detailed reference in our database. References found as HIPAA 164.312(a) were declare as "HIPAA 164" and so on. Also, no academic reference citation was kept. To understand the most influential documents, we calculate Betweeness Centrality based on average path length from references. This index[13], reflects the amount of influence exerted by a given node (standard reference) over the interactions between the other nodes in the relationship network Figure 4. This image shows the connections distribution and concentration for visual idea of mapping found. The density found on this graph is 0,6%. Guided by this rank table, an implementer can find documents that either has common features for HIS development (hence is heavily cited) or better describes them.

DISCUSSIONS

In our analysis proposition, word counting as part of statistical and text-processing steps can be influenced by different writing styles at some point and thus TF-IDF and VSM as well. For instance, the term EHR-S (electronic health record system) is commonly found also as "EHR system" along requirements texts. Some other influences during translation for processing the Portuguese (PT-BR) requirements portion added to study (SBIS requirements) was experience during database visual inspection. A few acronym needed to be corrected manually as well as other terms that were wrongly condensed to single word although expressing different meaning. A common example of those cases was the word "responsável" that after computer assisted translation (CAT) processing were all translated to "responsible". Despite CAT works semantically for some cases, in English writing there are a few other words that better express particular sense, like "accountable", "liable" or "bound". Therefore, the ideal condition for such analysis would be single language standard data-set.

Another major influence for VSM was the images and table contents not extracted using this method. For instance, the ASTM E2538 has almost 50% of its length composed by images and graphs.

There´s a significant difference between word content similarity and paraphrase analysis within documents. Although we have tested adding Wordnet to capture word sense matching instead on simple match, a more robust process possibly using specialized thesaurus may be required for precise results. Therefore, a similarity table cannot be seen as sole instrument for choosing which standard to discard based on text superposition as words arrange and context plays important role. The passphrase algorithm used is error prone when it comes to semantics association. The proposed synthesis may require a human analysis and interpretation for more accurate text representation.

Regarding the method approach, the main difficulty found was to set and configure Pdf-extract tool to properly extract the standards text contents. As each document has its own structure (even when comparing documents from same standardization organization), this settings required a particular configuration for each document.

Special attention must be carried while choosing a set of standards to support software construction in terms of legal requirements and also standard longevity. As we have seen, CCHIT ceased its operation by the end of the year. Systems based on that reference may now face the need for new basis research. The standardization neutrality not to bias the guidance content must also be assessed. It depends mostly on members and experts engaged and the publication flow transparency.

CONCLUSION

HIS construction based on relevant requirements considering the multitude of guidance resource available may require extensive research for proper selection. The existing content overlap and content interconnection among different documents complicate this process. Content analysis using text processing tools can ease the selection of proper documents to be used as base for system construction. In our experiment, we considered standards documents declared used by Brazilian implementers subjected to SBIS certification process, as we perform comparisons and relationships establishments to promote a better view of existing information security references.

ACKNOWLEDGMENT

We thank CAPES sponsorship that allowed proprietary standards acquisition for this study.

REFERÊNCIAS

1. Buckley FJ. An overview of the IEEE computer society standards process. Comput. Stand. Interfaces. 1987; 6(2):267-74.
2. Oksala S, Rutkowski A, Spring M, O'Donnell J. The structure of IT standardization. StandardView. 1996;4(1):9-22.
3. Leistner S. Avoiding surprises: some thoughts on standards. J IEEE Micro. 1998;18(3):25-32.
4. Johnson BC, Dunn DG, Hulett R. Seeking global harmony in standards. Ind Appl Mag IEEE. 2004;10(1):14-20.
5. Cyr TJSt. An overview of healthcare standards. Proceeding of the IEEE Southeastcon, 2013 Apr 4-7; Jacksonville, FL:IEEE; 2013. p. 1-5.
6. Hassan T, Baumgartner R. Intelligent text extraction from pdf documents. Vienna: IAWTIC; 2005.
7. Zhang D. Topic detection based on K-means. Proceeding of the 2011 International Conference on Electronics Communications and Control. 2011 Sep 9-11; Ningbo, China: IEEE; 2011. p. 2983–5.
8. Xinwu L. Research on text clustering algorithm based on K _ means and SOM. Proceeding of the International Symposium Intelligence Information Technology Workshops; 2008 Dez 21-22; Shangai, China: IEEE Computer Society; 2008. p.341-4.
9. Boban I, Mujkic A, Dugandzic I, Bijedic N, Hamulic I. Analysis of a social network. Proceeding of the 12th International Symposium on Applied Machine Intelligence and Informatics (SAMI); 2014 Jan 23-25; Herl'any, Slovakia:

IEEEXplore. p.129-32.
10. Lim VM-H, Wong SF, Lim TM. Automatic keyphrase extraction techniques: a review. Proceeding of the 2013 IEEE Symposium on Computers & Informatics; 2013 Apr 7-9; Langkawi, Malaysia. p.196-200.
11. Witten IH, Paynter GW, Frank E, Gutwin C, Nevill-Manning CG. KEA: practical automatic keyphrase extraction. Proceedings of the 4th ACM conference on Digital libraries. 1999 Aug 11-14; Berkeley, CA, USA: ACM; 1999. p.1-23.
12. Abraham SS, Idicula SM. Comparison of statistical and semantic similarity techniques for paraphrase identification. Proceeding of the 2012 International Conference on Data Science Engineering. 2012 Jul 18-20; Cochin, Kerala, Índia: IEEExplore. p.209-13.
13. 13.Ragland A, Yuan X, Jones B. Analyzing the relationship between CCHIT certification criteria and HIPAA. Proceeding of the IEEE Southeastcon, 2013 Apr 4-7; Jacksonville, FL: IEEE; 2013. p.1-5.