RI 11022 15 October 2011 Computer Science

IBM Research Report

Lessons From Existing Business Process Hierarchies for Building a New One for Traffic

Pietro Mazzoleni[1], Biplav Srivastava[2] and Srikanth Tamilselvam[3]

[1]IBM T. J. Watson Research Lab
P.O. Box 704, Yorktown Heights,
NY 10598, USA

[2]IBM Research Division
IBM India Research Lab
4, Block C, I.S.I.D. Campus, Vasant Kunj
New Delhi 110070, India.

[3]IBM Research Division
IBM India Research Lab
Block D4, Manyata Embassy Business Park
Nagawara, Outer Ring Road
Bangalore 560045, India.

Lessons From Existing Business Process Hierarchies for Building a New One for Traffic

Pietro Mazzoleni Biplav Srivastava Srikanth Tamilselvam

IBM Research - T.J.Watson and India Labs
Email: {pietro@us,sbiplav@in,srikanth.tamilselvam@in}.ibm.com

Abstract

A process hierarchy has the advantage of giving the business a holistic view at all times, promoting a common terminology for inter-organization communication, maintaining customer-centric focus, and enabling benchmarking with industry best-practices. Today, the same kind of efficiency is needed in public sector where government offers services like traffic, water, energy and public safety to citizens. Here, a government not only needs to communicate with its own agencies but also external stake holders like citizens, businesses and international organizations. However, despite its success in business sphere, business process hierarchies are not available in public areas like traffic.

In this paper, we first articulate design principles which process hierarchies should follow and then empirically analyze their adherence by the widely-used standard Process Classification Framework. Then, we use the lessons to create a new process hierarchy for traffic domain. For the purpose, we obtain the process content by automatically extracting it from public documents using harvesting tools and semi-automatically organizing it. We indirectly validate our approach in *automotive* industry where processes extracted from automotive terms match favorably with *known* relevant industries. The paper enhances the state-of-the-art in understanding of how existing process hierarchies are organized as well helps push its success to public sector (smarter cities) domains.

1 Introduction

Business Process Management (BPM) has popularized the discipline of using well-documented procedures, i.e., the business processes, at the center of all operational, management and support activities a business engages with its numerous partners, suppliers and employees[9]. To do so, a business process framework is needed.

17^{th} *International Conference on Management of Data COMAD 2011, Bangalore, India, December 19–21, 2011*

A business process framework is a "high-level, industry-neutral, enterprise process model that allows organizations to see their business processes from a cross-company viewpoint" [3]. Usually the framework is represented as a hierarchy and hence also called a Business Process Hierarchy (BPH). APQC's (American Productivity And Quality Center) Process Classification Framework (PCF) is an example of a widely-used hierarchy, also referred as map in the rest of the paper, which organizes cross-industry and industry-specific business processes. A process hierarchy has the advantage of giving the business a holistic view at all times, promoting a common terminology for inter-organization communication, maintaining customer-centric focus, and enabling benchmarking with industry best-practices.

Today, the same kind of efficiency is needed in public sector where government offers services like traffic, water, energy and public safety to citizens. Here, a government not only needs to communicate with its own agencies but also with external stake holders like citizens, businesses and international organizations. But much of the information sharing today is informal, ad-hoc and ambiguous. Consider traffic. There is no common terminology used to communicate across different parties or benchmarks to compare current and desired performances.

However, despite its success in business sphere, business processes hierarchies are not available in public areas like traffic. We want to build a business process hierarchy for traffic but not by the traditional approach of assembling it exclusively by a committee of experts. Instead, we propose a novel method where we leverage public documents related to traffic management and use content extraction techniques to automatically get process content. Our goal is to semi-automatically generate a business process hierarchy which can be reconciled with the standard APQC maps to further promote collaboration between public sector and private industries.

However, reconciling process content in BPHs needs understanding of how they should be organized. We first articulate the results of the work we did interviewing subject matter experts from multiple industries as well APQC to identify guiding principles that process hierarchies should follow. We then report on empirical analysis we conducted to verify the adherence of APQC to these principles. Then, we present our method to semi-automatically create a new process hierarchy for traffic domain. It was difficult to validate the quality of extracted traffic hierarchy; so we did a similar extraction for automotive industry and compared them with the known APQC hierarchies and found expected overlap with automotive and related industries. We then conclude with related work and pointers for future work. To the best of our knowledge, this is the first work on the topic. Furthermore, the feasibility of the new method shows that we can use it to rapidly build more hierarchies for other public sector services.

Our contributions are the following:

- Propose design principles for organizing process hierarchies and analyze standard APQC's BPHs for adherence to design principles.

- Present a solution to semi-automatically extract BPHs from generic documentation.

- Create a BPH for traffic using our solution.

- Validate the approach indirectly by extracting and comparing process hierarchies in a known industry.

2 Preliminaries

In this section, we first give background on APQC and then further motivate the problem of organizing traffic-related process content in hierarchies.

2.1 APQC Background

APQC's Process Classification Framework is a standard for terminology on process definitions and measures for benchmarking [3]. The development of APQC taxonomy started in 1992. More than 80 organizations have helped it evolve and it is now being used by thousands of organizations worldwide. Today, the BPH is the worlds most widely used process framework and allows organizations to speak a common language about functions, processes, and activities. The version we will refer to in the paper is 5.0.3. APQC is used as standard organizational principle for enterprise processes knowledge. Such organization is then used in combination with APQC open standards benchmarking database to compare the performance of the enterprise with organizations from any industry.

Unique Identifier	Source	Level 1 = Category	Level 2 = Process Group	Level 3 = Process	Level 4 = Activity
10009		8 Manage Financial Resources (10009)			
10728			8.1 Perform planning and management accounting (10728)		
10729			8.2 Perform revenue accounting (10729)		
10730			8.3 Perform general accounting and reporting (10730)		
10731			8.4 Manage fixed asset project accounting (10731)		
10732			8.5 Process payroll (10732)		
10733			8.6 Process accounts payable and expense reimbursements (10733)		
10756				8.6.1 Process accounts payable (AP) (10756)	
10869					8.6.1.1 Verify AP pay file with PO vendor master file (10869)
10870					8.6.1.2 Maintain/manage electronic commerce (10870)
10871					8.6.1.3 Audit invoices and key data in AP system (10871)
10872					8.6.1.4 Approve payments (10872)
10873					8.6.1.5 Process financial accruals and reversals (10873)
10874					8.6.1.6 Process taxes (10874)
10875					8.6.1.7 Research/resolve exceptions (10875)
10876					8.6.1.8 Process payments (10876)
10877					8.6.1.9 Respond to AP inquires (10877)
10878					8.6.1.10 Retain records (10878)
10879					8.6.1.11 Adjust accounting records (10879)
10757				8.6.2 Process expense reimbursements (10757)	
10880					8.6.2.1 Establish and communicate expense reimbursement policies and approval limits (10880)
10881					8.6.2.2 Capture and report relevant tax data (10881)
10882					8.6.2.3 Approve reimbursements and advances (10882)
10883					8.6.2.4 Process reimbursements and advances (10883)
10884					8.6.2.5 Manage personal accounts (10884)
10734			8.7 Manage treasury operations (10734)		
10735			8.8 Manage internal controls (10735)		
10736			8.9 Manage taxes (10736)		
10737			8.10 Manage international funds/consolidation (10737)		
10010		9 Acquire, Construct, and Manage Property (10010)			
11179		10 Manage Environmental Health and Safety (EHS) (11179)			

Figure 1: Example of APQC's BPH

Figure 1 shows an example of APQC process classification framework. APQC is organized into 5 hierarchical levels representing business functions at different levels of granularity. To convey the level of granularity, roughly put, a level 5 task can be done by an individual, a level 4 activity is done by one or a group of individuals, and a level 3 process is accomplished by an organization's group. Companies organize functions at level 1 and 2 in many ways depending on size, geography and other considerations. At each node, there is a number in the bracket referring to a unique identifier assigned in BPH.

APQC offers one, industry-neutral, hierarchy (also referred as *Cross-industry*) plus eleven industry-specific frameworks to describe processes (or process nuances) specific to the industry. Note that a company might be identified by more than one industry BPH, depending on their profile. As example, a large Media company will find processes in *Broadcasting* (for TV coverage), *Telecommunication* (for phone services) and *Customer products* (for new product development) BPHs.

Industry specific BPHs have all been created manually, by bringing Industry subjects matter experts along with companies intellectual capitals. Creating industry BPH is an extremely cumbersome and expensive activity, requiring endless number of iteration between members.

2.2 Building a Traffic BPH

The traffic of a city impacts all aspects of its citizens economic and personal activities. Here, government body managing a city's traffic not only needs to communicate with its own agencies but also external stake holders like citizens, businesses and even international organizations. A term like *traffic congestion* is used by traffic personnel, fire department, electrical utility, mayor office, citizens, civil contractors, IT companies implementing Intelligent Transportation Systems (ITS), etc. Although everyone is interested to know an accurate traffic outlook and to improve it, the common issues are:

- How do participants share common activities about the domain unambiguously?
- What kinds of information are really valuable?
- How to create the terminology in a practical, sustainable, manner?

A traffic business process hierarchy (BPH) could serve as the common terminology to communicate across different parties. Towards the aim of generating a *Traffic BPH*, we searched and collected numerous documents publicly available from municipalities all around the globe. Among the different document types we analyzed, we found *Request For Proposals(RFPs)* to be the most relevant for our work. RFPs are issued by cities or governments for building and managing traffic-related issues. Today, each US municipality, state, and central government independently publish their request from proposals. Last year, more than 27000 RFPs (from different authorities) have been issued in the US alone for Traffic-related issues [1]! When it comes to such a large number of proposals, how companies interested in participating can quickly find the RFPs relevant to them? How both Government and private companies can quickly identify which other solutions (processes) should be integrated with the ones described in the RFP? Which are the standard key performance indicators the proposed solution should meet? Similar to what happened in private Industry, we believe creating a standard BPH For traffic, aligned with the ones defined by APQC, will provide an answer to some of those issues.

3 BPH Design Principles and APQC Analysis

This section presents the guiding principles for generating well-formed BPHs. A BPH is meant to be consumed in the context of an industry (where cross-industry is a special industry) and furthermore, evolving relationships between industries should be evidenced in the overlap of BPHs. So, a BPH is well-formed when any industry expert can consume it without specific instructions. Moreover, a set of BPHs are well-formed when explicit guidance is available on how BPHs are structured. This second aspect is particularly important for our work as multiple industries (e.g., Automotive, Aerospace, Retails, Finance) may use (part of) the Traffic BPH to describe how their business processes interact with government agencies (municipalities) and other stake holders.

We now introduce 3 core BPH principles. They have been designed in collaboration with business process and industry subjects matter experts and by studying multiple BPHs created by IBM for describing customers' business processes. We will first list the principles and then explain them one-by-one in the context of APQC.

3.0.1 Principle 1:

A node is either defined in the Cross-industry BPH or it is specific to one single industry.

This first principle might seem obvious, but it is the cornerstone of a well-formed BPH. If not enforced, it will lead to the proliferation of nodes shared among only few industries, diminishing the value of the *Cross-industry* BPH. Moreover, as each BPH evolves independently of others, it would become infeasible to manage a *Cross-industry* BPH in which only nodes applicable to all industries are represented. To prevent such a problem, the first principle establishes that, if a node is shared by at least two industries, it should be defined in the *Cross-industry* BPH.

To validate, we ran an experiment to identify if APQC PCF adheres to this first principle. We took the eleven APQC maps (ten industries specific plus the *Cross-industry*) searching for nodes violating the principle. Not surprisingly, none of such nodes have been found: *in APQC, the only nodes shared across more than one industry are the ones listed in the cross industry BPH.*

3.0.2 Principle 2:

An industry-specific BPH should be self-contained, including all business processes relevant to that industry.

The first principle established nodes must be included in the *Cross-industry* BPH if they apply to at least two industries. On the other hand, not all *Cross-industry* nodes might be relevant to all industries. As example, APQC node 10134, "Gather current and historic order information" is relevant for most of the industries but not *Aerospace*. The second principle addresses this problem by requiring the industry BPH to include references to all relevant *Cross-industry* nodes. This approach makes each BPH self-contained, ready to be consumed, entity for companies operating in a specific sector.

This principle cannot be validated without an extensive knowledge of all industries. However, we studied how APQC industry BPHs reused *Cross-industry* nodes. The results of our experiment, depicted in Figure 2, shows how different industries have adopted

[1]Source: http://www.governmentbids.com

Industry Name	Tot Num Nodes	Industry BPH Nodes / Tot Nodes	Cross-Industry BPH nodes / Cross Industry nodes used in Industry BPH
AeroSpace	1535	0.2886	0.9613
Automotive	1479	0.3516	0.8442
Banking	1369	0.1812	0.9868
BroadCast	982	0.1538	0.7315
ConsumerProduct	1300	0.1869	0.9305
Cross industry	1136	-	-
Electric Utilities	1195	0.0569	0.9921
Petroleum(DownStream)	1209	0.0711	0.9886
Petroleum(UpStream)	871	0.1183	0.6761
Pharmaceutical	1257	0.113	0.9815
Telecommunication	1675	0.3713	0.9234

Figure 2: Distribution of *Cross-industry* nodes in AQPC Industry BPHs

different approaches when it came to re-using *Cross-industry* nodes. As example, *Electric Utility* BPH includes more than 99% of the *cross-industry* nodes while others, like *Broadcast*, only refer to 73% of the *cross-industry* map. Even the percentage of unique nodes in a BPH vary significatively, going from 7% of *Petrouleum (Downsteam)* to 37% for *Telecommunication*.

One possible explanation of this variance is that not all industries have been built at the same time, and some have undergone more revisions than others. However, as a general rule, we can confirm the trend of defining Industry BPH as combination of all relevant cross-industry nodes plus industry specific concepts.

3.0.3 Principle 3:

Industry-specific BPHs can reuse cross-industry nodes but they cannot alter the genealogy of Cross-industry BPH nodes.

This principle guides how industry specific BPHs should be built. An industry BPH should always start from the *Cross-industry* BPH, removing the nodes which do not apply and creating new nodes (and levels) for industry specific element. On the other hand, taking cross industry nodes (or levels) and moving those under a different parent should not be allowed.

The rationale for this principle is twofolds. First, it enables easy consumption by companies spanning multiple industries. As discussed in Section 2.1, companies operating across multiple sectors (like IBM, as example) will build their own BPHs by picking and choosing nodes from different BPHs. In this very common scenario, users will not be able to merge the BPHs if they do not understand why each industry has positioned the same *Cross-industry* nodes under different parents. The second value in enforcing principle this is the need to easily propagate changes done in the *Cross-industry* BPH to all industries referring such nodes.

Once again, we took APQC and run an experiment searching for scenarios violating the third principle. Surprisingly, we found some nodes, listed on Figure 3, which did not follow the principle.

As example, node 10737, *Manage international funds / consolidation* is defined in *Cross-industry* BPH as level three and at level 02 for *Telecommunication*. There could be multiple reasons for subject matter experts to define exceptions to the third principle. Recall from Section 2.1 that a level of PCF corresponds to the scale of activity in that industry and effort put ranging from business divisions for level 1 to individuals at level 5. We hypothesize that industry-specific experts are aware of this information for their particular industry but may not be able to reconcile it from other industries, latter being the key information which determines the level in *cross-industry* BPH. Hence, cross-industry levels should be carefully built and once in place, they should not be altered – the third principle.

4 Solution Blocks for a Traffic BPH

In this section, we describe the information extraction and the BPH comparison techniques we built to semi-automatically generate the *Traffic BPH*. The steps are illustrated in Figure 4.

The process begins with harvesting documents related to traffic management. The documents are processed with off-the-shelf content harvesting tools. To obtain the best results, the proposed information extraction solution considers both a process-ignorant (without any domain knowledge) scenario as well a scenario using APQC BPHs as pre-existing domain knowledge. The extracted content is then compared with existing maps and arranged into a new BPH based on the principles gleaned from analyzing APQC. Finally, subject matter experts use the "BPH Wizard tool" to review the generated map and refine it where necessarily.

In this work, the document types we consider are

APQCID	Cross Industry Node Name	Cross Industry Level	Industry Specific Name	Industry Spec Level
10342	4.5.5Manage returns; manage reverse logistics (10342)	3	Automotive	4
10364	4.5.5.1Authorize and process returns (10364)	4	Automotive	3
10066	2.1.4Confirm alignment of product/service concepts with business strategy (10066)	3	Consumer Product	4
10184	3.5.3Manage customer sales (10184)	3	Consumer Product	4
10737	8.10Manage international funds/consolidation (10737)	2	Telecommunication	3
10767	8.10.1Monitor international rates (10767)	3	Telecommunication	4
10768	8.10.2Manage transactions (10768)	3	Telecommunication	4
10769	8.10.3Monitor currency exposure/hedge currency (10769)	3	Telecommunication	4
10770	8.10.4Report results (10770)	3	Telecommunication	4
12341	2.1.7.3Design for manufacturability/assembly (12341)	4	Automotive	3
12651	5.5Train and manage customer service work force (NEW) (12651)	2	Automotive	3
11338	2.1.3.2Develop concepts (11338)	4	ConsumerProduct	3

Figure 3: Inconsistencies found by comparing Industry Specific and Cross-Industry nodes

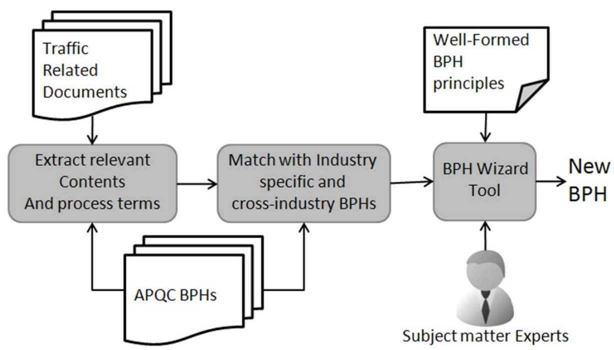

Figure 4: Steps in Creating the Traffic BPH.

non-restrictive and currently includes Request for Proposals (RFPs) that cities issue for traffic projects, public research papers on traffic sensing and analytics, and design documents of traffic solutions delivered to cities. Basing ourselves on documents allows us to identify important processes which figure in real-world efforts rather than those merely theoretically possible. Hence, the process of BPH creation becomes practical. Further, documents are very easily available and can be used by other researchers interested to this problem.

The next sections will describe the details of the steps we adopted to create the *Traffic BPH*.

4.1 BPH Terms Extraction

Content extraction, and the closely related area of information extraction, is a well researched sub-area of information management with many research prototypes and commercial tools. A good survey for information extraction techniques is [10]. Different approaches vary based on the type of documents considered (e.g., plain text, PDF, MS Word, Excel), the type of prior knowledge assumed (domain models), the supervision needed (e.g., supervised v/s unsupervised), and the type of concepts extracted (e.g., plain list v/s hierarchy).

As an illustration of the state-of-the-art, we work with IBM SPSS Text Analytics[11] which is a commercial tool that works with PDF, Word and other popular document formats. We will refer to it as *Tool1*. The tool performs unsupervised extraction and can generate concept hierarchies. During extraction, *Tool1* relies on linguistics-based text analysis. It first separates textual data from other content in the document and then identifies candidate concepts which can be made up of both single or multiple words. Then, it resolves synonyms and organizes concepts in a canonical form. The extracted concepts can be used to build a hierarchy. For this, *Tool1* uses a combination of linguistics and statistical methods to extract relationships among concepts. Further details on the techniques are available in its help documentation [11].

4.1.1 Iterative Approach.

Our objective is to extract the traffic BPH but in the process, also validate that the quality of hierarchy is good. We consider this in two parts: (a) generate a

BPH from traffic documents, and (b) for quality assessment, repeat the BPH extraction from automotive documents where APQC already has BPH. Then, in the next section, we discuss how the extracted BPHs are compared with existing APQC BPHs to confirm that the hierarchies are reasonable.

Table 1 summarizes the documents used (inputs) and processes extracted (output). A closer view is shown in Figure 5 where a subset of file-level statistics is given. The tool could effectively generate a rich set of concepts and arrange them in a detailed hierarchy of up to 6 levels. However, since we are not interested in general concepts but those related to activities (i.e., processes), we needed to do more. Content extraction tools allow filtering of results based on different rules. We started by using a simple rule to detect a process term: *if the phrase "process" is found in the name of a leaf concept, the concept along with its parents are deemed to refer to processes.*

However, the simple filtering rule tends to be too restrictive when compared to the activities found in APQC Maps. As example, *Analyze systems and technology* is a process relevant to Traffic which would get left out by the filter from the extracted processes. To address this limitation, we analyzed the terms in the existing APQC maps looking for terms that were most commonly used to define processes. Figure 6 shows the word Cloud created for APQC BPH Maps using Many Eyes technique[8]. The 5 most common terms are perform, strategy, customer, establish and define. We use them along with "process" as process indicators. That is also the filter reported in Table 1.

The concepts extracted for each document have been combined generating a total of 180 process related concepts for the complete traffic dataset. Figure7 shows a snapshot of the hierarchy extracted by *Tool1*. Many branches of the hierarchy are intuitive. For example, *fare collection* is a process which is part of *fare* management activity and related to *finance*. Others are syntactically accurate but not intuitive. For example, *process fleet* and *processing of probe* are *processors* but not necessarily related to *electronic equipment*. But given the evidence of the documents, the tool generated it.

To validate our approach, we ran a similar extraction on 40 *Automotive* documents. We found 538 process-related terms organized in a six level hierarchy. Figure 8 shows a snapshot of the hierarchy extracted. The extracted hierarchies are compared with existing APQC BPHs next.

4.2 Mapping Extracted Concepts with APQC BPHs

As next step, we compared the process terms extracted for traffic and automotive domains with the nodes in the eleven APQC hierarchies. The goal of the comparison is to identify how the *Traffic* BPH should be built as a combination of cross-industry and industry specific nodes, as well as validate it in the case of *Automotive* where we already have a BPH. Two nodes, s_i and s_j, are said to be equivalent if they have matching names. Suppose θ is a text similarity measure viz. edit-distance by Monge Elkan method[4], and ϕ is a threshold. The matching function between s_i and s_j is given below.

$$\begin{aligned}(s_i, s_j) &= Match, \theta_t(s_i, s_j) \succ \phi, \\ &\quad \bot, otherwise. \end{aligned} \tag{1}$$

Here, we compare the individual terms from the traffic/auto hierarchy with the terms of the APQC BPHs and record matches. We also record whether the matching BPH nodes are from cross-industry or industry-specific maps, as well as how many do not match. Note that our match definition entails reporting all possible matchings above a threshold. Hence, the # matches can be more than the # terms in either the extracted hierarchy or the BPHs when there are more than 1 possible matches. As part of generating the final BPH, an eventual match decision has to be made (by a human or rules) as discussed in Section 4.3.

4.2.1 Experiments.

Table 9 shows the results of matching the *Traffic* BPH consisting of 180 terms obtained with the expanded process filter with APQC BPHs. The first column lists the industries. The second column shows how many nodes from the APQC BPH matches the process terms in the extracted hierarchy. The third column shows it in percentage. The fourth column shows how many matches were found by considering only the portion of the BPH referring to *Cross-industry* nodes (fifth column has this in %). The sixth column shows how many matches were found against industry specific nodes (seventh column has this in %). As example, the APQC *Automotive* BPH contains 1479 contains, 520 of which are specific to the industry (see Table 2). Table 9 shows that 224 concepts from traffic match the industry BPH. The percentage in column seven is computed by dividing 224 (Matches) against the base for comparison (520). The last two columns show how many traffic terms did not match that BPH in numbers and %, respectively.

Many observations can be made from the results:

- In the ideal case, the number of matches should be equal to the number of extracted concepts. However, since matches are approximate, many more potential matches are found.
- Recall from Table 2 that a high proportion of cross-industry nodes show up in the industry-specific BPHs. The results give a breakup of the

S.No.	Traffic	Automotive
#Documents	18	40
#Levels	6	6
#Process Terms	180	538

Table 1: Statistics on documents used in experiments and final processes extracted.

Region	Year	Document Type	Pages	# Concepts
Surat, Gujarat India	2010	RFP	79	1674
Ahmedabad, Gujarat India	2010	Tender for ITS	79	1196
Michigan Dept, of Tranportation USA	2011	RFP	19	573
City of SeaSide, USA	2010	RFP	6	729
City of New Rochelle, USA	2007	RFP	4	151
New Delhi, India	2010	Research Report	5	250
Baltimore Metropolitan Council, USA	2011	RFP	11	230
City of Urbana, Illinois, USA	2010	RFP	14	306
USA	2010	Evaluation ppt	18	288
USA	2010	Catalog, doc	23	173

Figure 5: Detailed statistics of a subset of documents used in experiments.

matches in columns six and eight. *Telecommunication, Aerospace, Automotive* and *Consumer Products* have the top-most number of industry-specific nodes matching. Closer inspection reveals that this is reasonable because many activities in air traffic management are similar to traffic management, and so on.

- More than 75% of the traffic nodes are not matching any of the known BPHs. Although traffic related, they are either not covered in current BPHs (hence refer to novel processes) or too specific to be generalized and made part of a BPH.

To check whether the results were sensitive to the filtering terms, we ran another experiment on a subset of documents (shown in Figure 5) but considered only *process* as the process filter. We got 99 process terms related to traffic. The top matching industries for the extracted traffic processes remained the same (*Telecommunication, Aerospace, Automotive* and *Consumer Products*) but *Aerospace* was the top-most instead of *Telecommunication* with more terms related to traffic management. The extracted processes were more sensitive to document samples.

Table 10 has the similar results for extracted automotive process terms. Some observations from the results are:

- The extracted terms have a high match with industry-specific nodes of *Automotive* BPH including those which match only with this industry.
- Interestingly, the highest match of the extracted hierarchy is with nodes from *Telecommunication* industry. This is due to common terms related to performance measurements.
- More than 85% of the automotive terms are not matching any of the known BPHs. Although automotive related, they are too specific to be generalized and made part of a BPH.

The automotive result validates that the approach can generate process hierarchies consistent with what is already known. Further, the traffic results show that the generated hierarchy in a new domain can be reasonable. Together, they support the presented approach as a promising way to obtain process hierarchies in new domains like public sector.

However, the above results are only a start. They are a function of the sample of documents used and the similarity function used for comparison. We have tried to use a large set of documents from different regions of the world relating to all aspects of traffic and automotive operations. Still, there is much room for future investigation.

4.3 Refining Traffic BPH

Content extraction and BPH comparisons techniques are two important steps toward the automatic generation of a new BPH. However, for consistency, generality and other reasons, not all steps can be automate. In fact, one would want subjects matter experts to review the generated BPH and refine it. This sections describes the types of refinements required to build a well-formed BPH based on the principles listed in Section 3.

The BPHs comparison process clusters each of the extracted process nodes into one of the following scenarios:

Figure 6: Word Cloud considering all APQC BPH terms

1. The extracted node matches a cross-industry node. In this case, the node should be defined in the new BPH as a reference to the cross-industry node.
2. The extracted node does not match any pre-existing nodes. In this case, the node should be added as industry specific process with suitable adjustment to match APQC's naming convention.
3. The extracted node matches a node in a different industry BPH. In this last case, the node cannot be shared among two industries, as it violates our well-formed BPH principle 1. Based on how closely related are the two concepts, the expert will either opt to decouple the concepts and add it as different industry specific node, or to maintain the relation and move the node to the cross-industry BPH. This second option is preferred as promote interoperability among industries.

The three scenarios provide guidances on how subject matter experts should refine each generated node. However, those scenarios do not take into consideration the hierarchical relations among nodes. In fact, consider as example a node, called A, which does not match any pre-existing node (scenario 1), but one of its children nodes, called B, is part of cross-industry BPH (scenario 2 above). In this scenario, B cannot be added as reference to the cross-industry node A as it violates BPH principle 4 (cross-industry nodes cannot be shuffled to different parents).

To address all possible scenarios (each node can in fact have multiple ancestors and descendants) we designed a *BPH Wizard Tool* which takes in input the generated BPH and guides experts in creating a well-formed BPH. The *BPH Wizard Tool* starts with a post-tree traversal of the BPH tree searching for branches containing nodes all belonging to scenario tree above (nodes match existing industry-specific nodes). Those branches are addressed first as behave as single nodes. The experts will be asked to decide if the nodes/branches should be industry specific or references to cross-industry nodes. When completed, the wizard continues with an in-order traversal of the BPH searching for branches including both cross-industry and industry-specific nodes. In-order tree traversal is done because high level nodes, which represent broader industry processes, should be addressed first. A minimum impact approach is suggested to resolve these scenarios.

Once all branches are resolved, the last step in creating well-formed BPH is making sure the new map is self contained. Once again, we make use of the results of the concept mapping extraction process. Specifically, we search for cross-industry nodes which have been included in most of the industries but not found in the generated BPH. If found, the wizard will suggest to the expert to either look at those nodes manually or to repeat the content extraction providing any additional industry specific documentation as input.

5 Related Work

The importance of process classification frameworks for realizing BPM potential is well known[1]. As public services become increasingly IT-enabled, BPHs will become important in this area as well.

There is a rich literature on process design, much of which is centered around streamlining e-commerce. In [7], the authors propose (financial) value creation as the basis for organizing cross-organizational business processes. They consider a hierarchical structure for processes and they use levels to distinguish between *Primary*, *Support* and *Management Processes*, depending on the role the process plays in satisfying the customer need. However, this is not readily applicable in public sector where money is not the explicit basis for cooperation. In another work[12], ontology is proposed for business automation. The domain knowledge of processes is formalized in a process ontology that enables agents communication, sharing of information and reuse of capabilities. However, a domain ontology does not exist for the domains we are considering.

Away from e-commerce, [2] proposes a framework to manage an organizations process knowledge along

Level 1	Level 2	Level 3	Level 4	Level 5
finance				
	fare			
		fare collection		
			fare collection process	
		security		
			bid evaluation process	
			bid opening process	
transportation				
	vehicles			
		probe vehicles		
			processing vehicle probe	
electronic equipment				
	processors			
		process fleet		
		current dot collection processes		
		processing efforts		
		processing of the intellidrivesm		
		processing of probe		
		processing system		
			electronic equipment	
				traditional systems engineering process
				systems engineering process
business management				

Figure 7: Snapshot of our *Traffic* BPH process terms and categories extracted with *Tool1*. Filter=≺process,perform, strategy, customer, establish, define≻.

seven dimensions: structure, personnel and coordination, performance and tools, discourse, results, quality and implications. Once process hierarchies are in place, efforts can be designed for process improvement[5]. They are applicable to our work as well.

In this paper, we used documents to discover processes. There is a rich literature on mining processes from executable systems[6]. However, although it is not easily applicable in our case because IT-driven traffic systems (called Intelligent Traffic Systems) are commonly not in place, and where available, they are considered core city infrastructure and thus restricted for data sharing with others.

6 Conclusion

In this paper, we considered the problem of creating a BPH in the sub-domain of public services, specifically traffic, in an efficient manner while reusing existing APQC BPHs. We proposed three design principles for well-formed BPHs and evaluated those existing maps. Then, we use the lessons and existing APQC maps to create a new process hierarchy for traffic domain based on concepts gleaned from document using content extraction methods. The paper enhances the state-of-the-art in understanding of existing process hierarchy organization as well helps push its success to public sector (smarter cities) domains.

References

[1] C. Aitken, C. Stephenson, and R. Brinkworth. Process classification frameworks. In *Handbook on Business Process Management 2*, International Handbooks on Information Systems, pages 73–92. Springer Berlin Heidelberg, 2010.

[2] C. S. Amaravadi and I. Lee. The dimensions of process knowledge. In *Knowledge and Process Management, 12: 6576*, 2005.

[3] APQC. Apqc process classification framework (pcf)., 2008. available at http://www.apqc.org/.

[4] W. W. Cohen, P. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proc. Information Extraction on the Web Workshop (IIWeb)*, 2003.

[5] P. Dalmaris, E. Tsui, B. Hall, and B. Smith. A framework for the improvement of knowledge-intensive business processes. In *Business Process Management Journal, Vol. 13 Iss: 2, pp.279 - 305*, 2007.

[6] W. V. der Aalst, B. V. Dongen, J. Herbst, L. Maruster, G. Schimm, and A. Weijters. Workflow mining: A survey of issues and approaches. In *Data and Knowledge Engineering, 47 (2), 237-267*, 2003.

Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
finance					
	key performance indicators				
	credit				
		customers credit application			
	investment				
		future			
			future of the establishment		
			future customer		
	liquidity				
		financial performance			
	leasing				
		leasing customers			
	financing				
		finance customer relationship			
automotive					
	automotive customer lifecycle				
	automotive repair				
manufacturing					
	established manufacturing operations				
	factory				
		assembly plants			
			established assembly plants		
			customers assembly plant		

Figure 8: Snapshot of our *Automotive* BPH process terms and categories extracted with *Tool1*. Filter=$\prec$perform, strategy, customer, establish, define$\succ$.

[7] J. Gordijn and R. J. Wieringa. A value-oriented approach to e-business process design. In *CAiSE, LNCS 2681, pp. 390403*, 2003.

[8] IBM. Many eyes. Website, 2011. Available at: http://www-958.ibm.com/software/data/cognos/manyeyes/.

[9] N. Melo. e-business processes and e-business process modelling: A state-of-art overview. *Intl Jour. of Services Tech. and Mgmt*, 11(3):pp. 293–322, 02 2009.

[10] S. Sarawagi. Information extaction. In *Foundations and Trends in Databases, Vol. 1, No. 3, Pg. 261 to 377*, 2007.

[11] I. SPSS. Ibm spss text analytics. Website, 2011. Available at: http://www.spss.com/software/statistics/text-analytics-for-surveys/.

[12] B. Wu, L. Li, and Y. Yang. Ontological approach towards e-business process automation. *E-Business Engineering, IEEE International Conference on*, 0:154–161, 2006.

APQC Maps	Tot. Matched concepts		Matches against Cross Industry Nodes		Matches for Industry Specific Nodes		UnMatched concepts	
	Num	Perc	Num	Perc	Num	Perc	Num	Perc
AeroSpace	614	40%	450	41%	164	37%	141	78%
Automotive	614	41%	390	41%	224	43%	147	82%
Banking	543	39%	453	40%	90	36%	142	79%
BroadCast	375	38%	307	37%	68	45%	146	81%
ConsumerProduct	609	46%	395	37%	214	88%	142	79%
CrossIndustry	491	43%	491	43%	0	N/A	148	82%
ElectricUtilities	528	44%	454	40%	74	109%	143	79%
Petroleum(DownStream)	481	39%	448	40%	33	38%	143	79%
Petroleum(UpStream)	359	41%	315	41%	44	43%	153	85%
Pharmaceutical	502	39%	451	40%	51	36%	142	79%
Telecommunication	761	45%	458	43%	303	49%	146	81%

Figure 9: Results of matching extracted traffic terms with BPHs [ϕ=0.9, Filter=$\prec$process, perform, strategy, customer, establish, define$\succ$]

.

APQC Maps	Tot. Matched concepts		Matches against Cross Industry Nodes		Matches for Industry Specific Nodes		UnMatched concepts	
	Num	Perc	Num	Perc	Num	Perc	Num	Perc
AeroSpace	668	43%	495	45%	173	39%	457	84%
Automotive	727	49%	452	47%	275	53%	457	84%
Banking	625	45%	519	46%	106	43%	456	84%
BroadCast	396	40%	338	41%	58	38%	472	87%
ConsumerProduct	658	50%	459	43%	199	82%	457	84%
CrossIndustry	557	49%	557	49%	0	N/A	464	86%
ElectricUtilities	575	48%	511	45%	64	94%	455	84%
Petroleum(DownStream)	571	47%	510	45%	61	71%	456	84%
Petroleum(UpStream)	361	41%	335	44%	26	25%	481	89%
Pharmaceutical	580	46%	514	46%	66	46%	455	84%
Telecommunication	780	46%	495	47%	285	46%	460	85%

Figure 10: Results of matching extracted Automotive industry terms with BPHs [ϕ=0.9, Filter=$\prec$perform, strategy, customer, establish, define$\succ$]

.