Towards the Information Technology Usage for E-Government Portal Assessment based on Web Data Extraction Techniques

Andrii Kopp [1] and Oleksandr Chornenkyi [2]

[1] *National Technical University "Kharkiv Polytechnic Institute", Kyrpychova str. 2, Kharkiv, 61002, Ukraine*
[2] *V.N. Karazin Kharkiv National University, Svobody sq. 4, Kharkiv, 61022, Ukraine*

Abstract

Today, interdisciplinary studies in computer science and social sciences, including political science, are inevitable due to the need to work with web-based sources to gain valuable insights, process large amounts of data, and apply various data analysis techniques. Web data extraction or web scraping is important for social and political studies when it is necessary to retrieve data arrays from a website for future analytical processing. Such automatic data collection and processing is a promising interdisciplinary field for social scientists and computer scientists. Therefore, this study aims to improve e-government web portal evaluation processes by proposing a corresponding information technology based on web data extraction techniques. The software implementation of the proposed technology is based on Python and Power BI for computation and visualization, respectively. The proposed toolkit was used to analyze the e-government web portals of two countries selected on the basis of their high e-government development index, the obtained results of prevailing services on each of citizen portals were analyzed and discussed, and the corresponding conclusions were made.

Keywords [1]

E-Government Web Portal, Citizen Portal, Information Technology, Web Data Extraction.

1. Introduction

1.1. Motivation

Nowadays, the rapid evolution of computer technologies changes scientific approaches to modern issues and provides ways for creating novel and enhancing existing research methods. It is especially considerable for applied science wherein computational technique implementation accelerates the complex applied problem solution requiring large volumes of calculations. Social sciences, which have historical relations with philosophy, have a peculiar wide range of research methods. However, social sciences are also in a transformation state and increasingly using computing technology for research problem solving, which has led to the emergence of computational social sciences. Initially, computational social sciences were associated with agent-based modeling for the simulation of the behavior of an individual or social group under certain conditions. Nevertheless, the Internet spreading, social networks and online platforms popularity increasing within the growth of numbers of Internet users provoked a new large stream of digital data which has become a valuable source of information for social sciences researchers and has led to the expansion of the concept of "computational social science" [1]. Although earlier researchers have argued that digital data analysis-based computational social science has developed slowly [2], more recent studies show that in recent times increased the interest of social sciences scholars in using computational techniques for research [1].

1.2. Related Work

The use of information technologies for political science research is not as new as it may seem, and began in the second half of the 20th century. The first experiments using computers were aimed at

Information Technology and Implementation (IT&I-2023), November 20-21, 2023, Kyiv, Ukraine
EMAIL: kopp93@gmail.com (A. Kopp); chornenkyi.o.o@gmail.com (O. Chornenkyi)
ORCID: 0000-0002-3189-5623 (A. Kopp); 0009-0001-9479-1776 (O. Chornenkyi)


trying to predict election results. Typically, all studies were based on the use of agent-based modeling in different form, using classical theories of political interactions [3]. The further evolution of the Internet, increasing the power and availability of computer technology provided new research fields for political scientists and tools for expanding methodology. Political science methodology expansion led to the fact that in addition to agent-based modeling, political scientists more often were beginning to use methods related to big data analysis [4].

Scholars in social and political science address "big data" as a broad concept that includes any digital elements left by users or organizations on the Internet that can be read by information technologies [5]. In the computer science field, big data is usually associated with so-called "5Vs" used to describe its characteristics (value, variety, velocity, veracity, and volume) [6].

Today, for social sciences fields and particularly for political science, when working with big data, it is important to use web scraping tools, which is the automatic extraction of data from websites for further analytical processing [7]. For political science, this approach can be valuable for defining features of how political parties or government agencies use their websites [5] and for researching local politics through the mining and analysis of unstructured data from the websites of local government institutions [8]. Some researchers conclude and we agree that the use of automated data mining in social sciences opens a new way for cooperation between social and computer science researchers [6].

It should be stressed that political science must be in the continuous dynamic movement condition and must permanently react to political changes in the modern world. Today, the policies of many countries aim at the formation of an inclusive information society, which includes widespread digital transformation. Governments create and support open e-government web portals, which aim to facilitate citizens' access to government information and improve the process of providing government services to citizens. However, it should be noted that the politics of different states differ from each other and may have various accents. Under such circumstances, it may be interesting how the policies of different states influence their e-government web portals.

Thus, we propose an information technology that can help researchers to explore services provided on government web portals, which can be useful for further analysis of different countries' policies. It aims to improve e-government web portal evaluation processes by using web harvesting techniques.

Therefore, this study is expected to answer the following research questions:

- What reference model can be used to evaluate the e-government web portal?
- What algorithms can be used to process and harvest the desired e-government web portal data?
- How can the extracted data be quantitatively evaluated to compare the policies of different countries and define the prevailing citizen services?

2. Materials and Methods

2.1. E-Government Web Portal Services Model

Let us formally describe the set of services that the e-government web portal is expected to provide:

$$eGS = \{eGS_1, eGS_2, \dots, eGS_n\}. \quad (1)$$

Here n is the number of services $eGS_1, eGS_2, \dots, eGS_n$ the e-government web portal is expected to provide, $i = \overline{1,n}$.

Moreover, for each of the e-government web portal services eGS_i, $i = \overline{1,n}$ we propose to define the set of keywords W_i, $i = \overline{1,n}$ which completely describes the mentioned service:

$$\delta: eGS_i \to W_i = \{w_{i1}, w_{i2}, \dots, w_{im_i}\}. \quad (2)$$

Here m_i is the number of synonymic keywords $w_{i1}, w_{i2}, \dots, w_{im_i}$ in W_i, $i = \overline{1,n}$ defined for i-th e-government web portal service eGS_i, $j = \overline{1,m_i}$.

Hence, the formal definition of E-Government Web Portal Services (EGWPS) model can be formulated as given below:

$$eGWPS = \langle eGS, \delta, W \rangle. \quad (3)$$

Here W is the set of keyword sets mapped to each of the e-government web portal services, $W = \{W_1, W_2, \dots, W_n\}$. Let us graphically illustrate in Fig. 1 the proposed e-government web portal services

model. Fig. 1 demonstrates the set of expected services and their keywords used to detect such services on the e-government web portals under assessment. Using the proposed EGWPS model (Fig. 1), we propose to find the "distance" between the e-government web portal under assessment and the so-called "perfect" e-government web portal (in terms of its contents) described by this model.

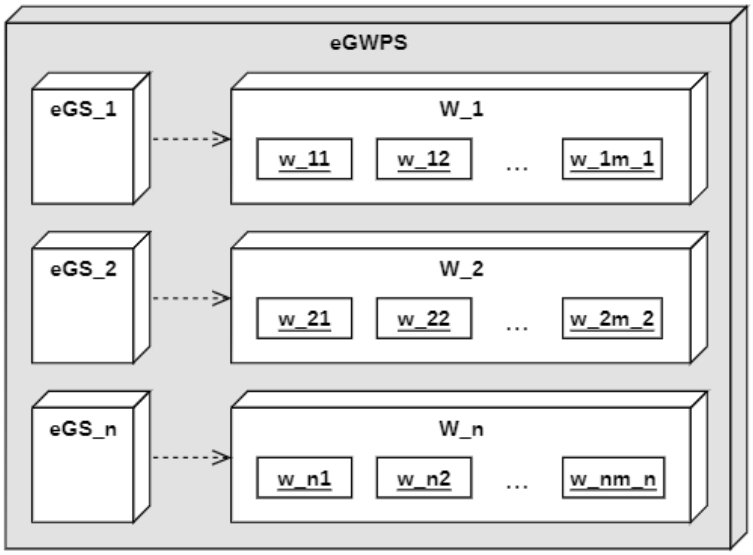

Figure 1: Proposed e-government web portal services model

Therefore, we propose to use the web data extraction (or web scraping, web harvesting etc.) technique to automatically explore and assess e-government web portals against the proposed EGWPS model (Fig. 1).

2.2. Web Data Extraction Algorithm

Fig. 2 below illustrates the e-government web portal data meta-model given using the UML (Unified Modeling Language) [9] class diagram.

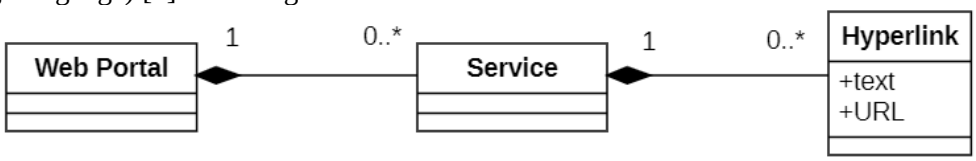

Figure 2: Meta-model of the e-government web portal data

The e-government services data extracted from HTML (Hyper Text Markup Language) pages of a corresponding web portal is represented as the set of HTML hyperlink tags [10]:

$$H = \{h_1, h_2, \dots, h_p\}. \quad (4)$$

Here p is the number of hyperlinks $h_1, h_2, \dots, h_p$ located on e-government web portal HTML pages, $k = \overline{1,p}$.

Each hyperlink tag includes the text and URL (Unified Resource Locator) as it is demonstrated in the meta-model diagram in Fig. 2.

The set of e-government web portal services eGS extracted from the respective HTML pages is represented as follows:

$$S = \{S_1, S_2, \dots, S_n\}. \quad (5)$$

Each service S_i, $i = \overline{1,n}$ is expected to be digitally implemented by one or multiple hyperlinks located on the e-government web portal pages:

$$S_i = H_{S_i} \subseteq H. \quad (6)$$

Here H_{S_i} is the sub-set of hyperlinks extracted from the e-government web portal that implement i-th service detected in the e-government web portal S_i, $i = \overline{1,n}$. Therefore, to detect services provided

by the e-government web portal using the proposed EGWPS model (Fig. 1) and the meta-model (Fig. 2), the following algorithm should be used:

Given: set of extracted e-government web portal hyperlinks H
EGWPS model $\langle eGS, \delta, W \rangle$
empty set of detected e-government web portal services S
for each eGS_i in eGS:
  for each h_k in H:
    $W_i = \delta(eGS_i)$
    for each w_{ij} in W_i:
      if w_{ij} is a substring of h_k text:
        $S_i \leftarrow h_k$
    end
  end
end

The input set H of the e-government web portal hyperlinks can be extracted using web scraping tools in Python or other programming languages.

The output set S basically represents the instances of Service class defined in the proposed meta-model (Fig. 2). Furthermore, each service S_i, $i = \overline{1,n}$ has multiple hyperlinks that belong to H.

Fig. 3 graphically illustrates the proposed algorithm using the UML activity diagram [9].

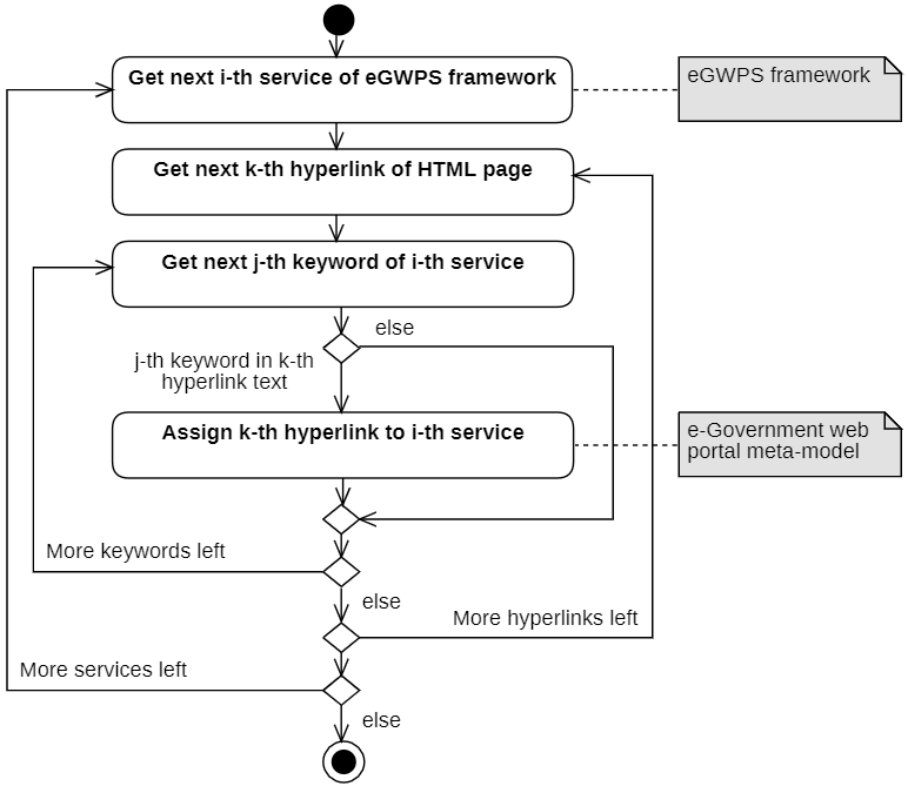

Figure 3: Proposed algorithm for e-government services data extraction from a web portal

2.3. E-Government Portal Assessment Metrics

Finally, we propose the following metrics to assess the e-government web portal in terms of detected services. The following metric allows to find the number of services detected in the e-government web portal under assessment:

$$SD = |\{S_i \in S, S_i \neq \emptyset\}|. \quad (7)$$

The following metric allows to find the "service richness" calculated as the relative number of services of the e-government web portal under assessment in comparison to the reference EGWPS model (Fig. 1) [11]:

$$SR = \frac{1}{n} SD. \quad (8)$$

The following metric allows to find the "relative cardinality" of the particular service calculated as the relative number of hyperlinks used to implement the i-th service detected in the e-government web portal under assessment in comparison to the maximum possible number of hyperlinks used in the same web portal for a certain service [12]:

$$SC_i = \begin{cases} \frac{1}{\max\limits_{i=1,n} |S_i|} |S_i|, & \max\limits_{i=1,n} |S_i| > 0 \\ 0, & \max\limits_{i=1,n} |S_i| = 0 \end{cases} \quad (9)$$

The following metric allows to find the total "service balance" to assess the balance of hyperlinks related to services detected in the e-government web portal under assessment:

$$SB = \frac{1}{n} \sum_{i=1}^{n} SC_i. \quad (10)$$

Using the following algorithm, it is possible to evaluate an e-government web portal against the EGWPS model. Thus, as a reference model, we can use the experience and best practices of the most advanced e-government web portals, define the set of services eGS a portal is expected to provide, and the keywords W to detect such services in corresponding HTML web pages.

Fig. 4 graphically illustrates the proposed algorithm using the UML activity diagram [9].

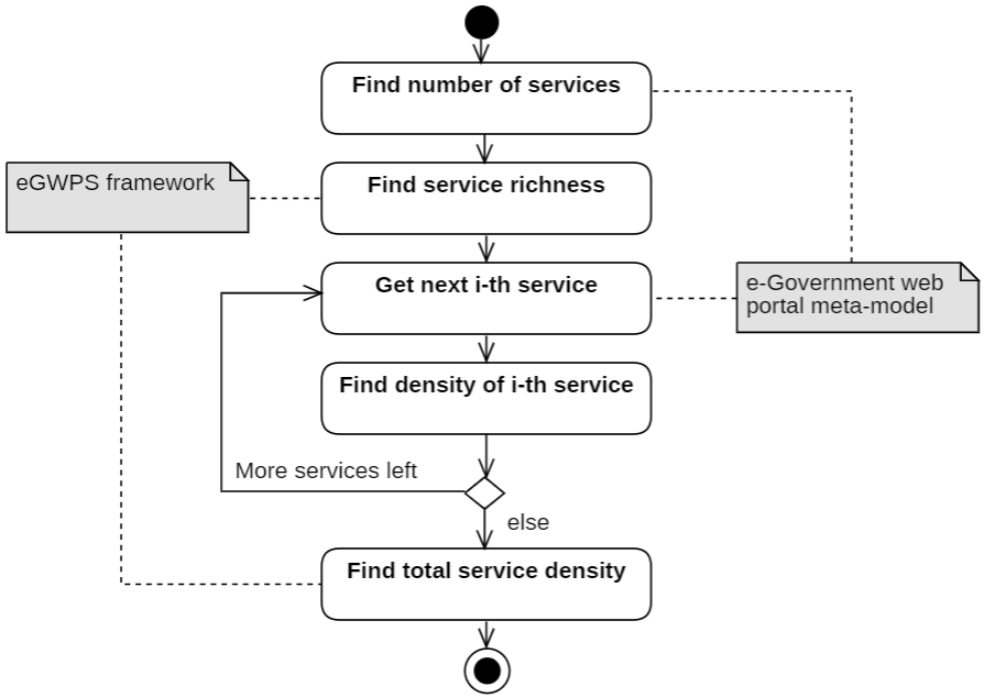

Figure 4: Proposed algorithm for e-government services evaluation

2.4. Information Technology for E-Government Portal Assessment

Finally, the information technology for e-government web portal assessment can be formally described using the following tuple:

$$eGWPAIT = \langle eGWPS, S, AM \rangle. \quad (11)$$

Here AM is the algorithmic model, which includes the proposed algorithms (Fig. 3 and Fig. 4) [13]:

$$AM = (A = \{A_1, A_2\}, R \subset A \times A). \quad (12)$$

Here A is the set of algorithms, where A_1 is the data extraction algorithm and A_2 is the evaluation algorithm; R describes the interconnections between the proposed algorithms when used to assess an e-

government web portal. Selected e-government web portals will be analyzed using the proposed information technology implemented using Python, in-build packages, and third-party libraries:

- "urllib" – used the "request" module to open and work with URLs [14];
- "re" – for regular expressions operations to parse web pages [15];
- "json" – to save results as JSON (JavaScript Object Notation) [16];
- "bs4" – used the "Beautiful Soup" library to scrape information from web pages of citizen portals [17].

Fig. 5 below demonstrates the Data Flow Diagram (DFD) [18] of the data processing workflow implemented by the proposed information technology.

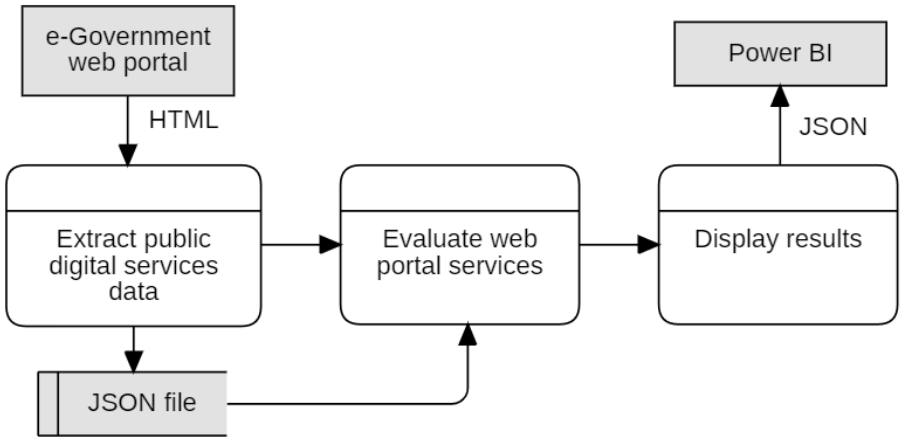

Figure 5: Data processing workflow for e-government portal assessment

According to Fig. 5, obtained results are displayed using Power BI – a high-performance Business Intelligence (BI) tool for advanced data visualization and data-driven decision making [19].

Fig. 5 illustrates the hands-on usage of the proposed information technology.

3. Results and Discussion

3.1. E-Government Portal Data Extraction: Citizens Viewpoint

Let us form the reference EGWPS model considering the Integrated Architecture Framework for E-Government (IAFEG) shown in Fig. 6 [20]. In this study, we focus on the "Social Sub-system" layer of this framework, in particular – on its "Citizens" perspective [20]. The citizens' viewpoint according to IAFEG [20] includes the following services (or topics) expected from an e-government web-portal:

- taxation;
- education;
- health;
- immigration;
- employment.

The IAFEG-based set of services the e-government web portals are expected to provide eGS (from the citizens' viewpoint of the IAFEG "Social Sub-system" layer) and the keywords W used to describe each of the services on HTML web pages are given in Table 1.

According to the "UN E-Government Knowledgebase" and its UN (United Nations) E-Government Survey 2022, top five countries with the highest E-Government Development Index are Denmark (0.9717), Finland (0.9533), Republic of Korea (0.9529), New Zealand (0.9432), and Iceland (0.9410).

Denmark citizen portal "Life in Denmark.dk" is shown in Fig. 7 [21]. The "Life in Denmark.dk" portal offers topics related to immigration, housing, working, family and children, money and taxation, education, healthcare, travel and transportation, pension, rights, leisure and networking, as well as stand-alone digital services (Fig. 7) [21]. Table 2 shows hyperlinks detected on the "Life in Denmark.dk" citizen portal according to IAFEG taxation, education, health, immigration, and employment services [20]. The "Suomi.fi" portal offers similar topics to "Life in Denmark.dk". These topics are connected to family, social security, healthcare, education, working, housing, rights and

obligations, finances and taxation, moving and travelling (Fig. 8) [22]. Table 3 shows hyperlinks detected on the "Suomi.fi" citizen portal according to IAFEG [20].

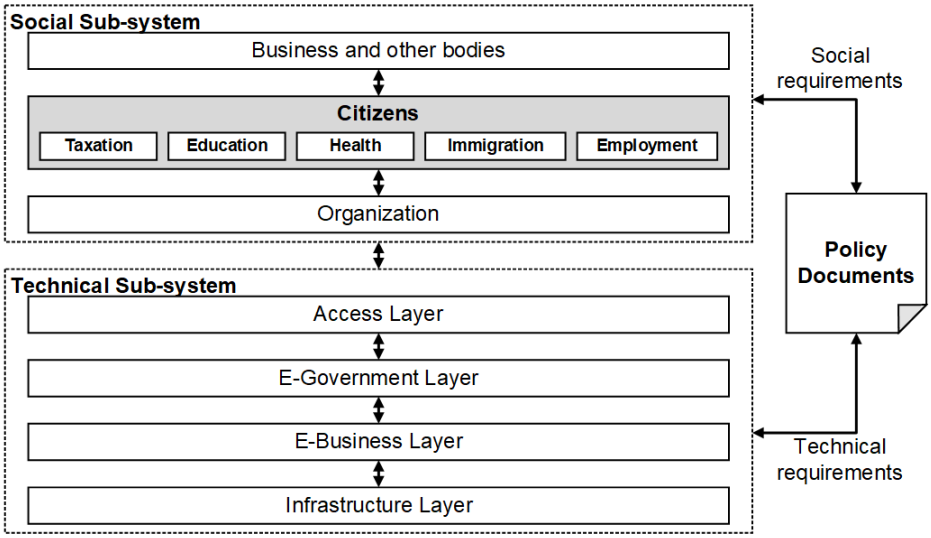

Figure 6: Integrated Architecture Framework for E-Government [20]

Table 1

Proposed EGWPS model contents based on IAFEG [20]

Services	Keywords
Taxation	tax, finance, income, money, debt, credit
Education	education, school, study, child, training, student
Health	health, insurance, care, sick, medical, funeral
Immigration	immigration, citizen, travel, visa, residence, international
Employment	employment, work, job, business, license, certification

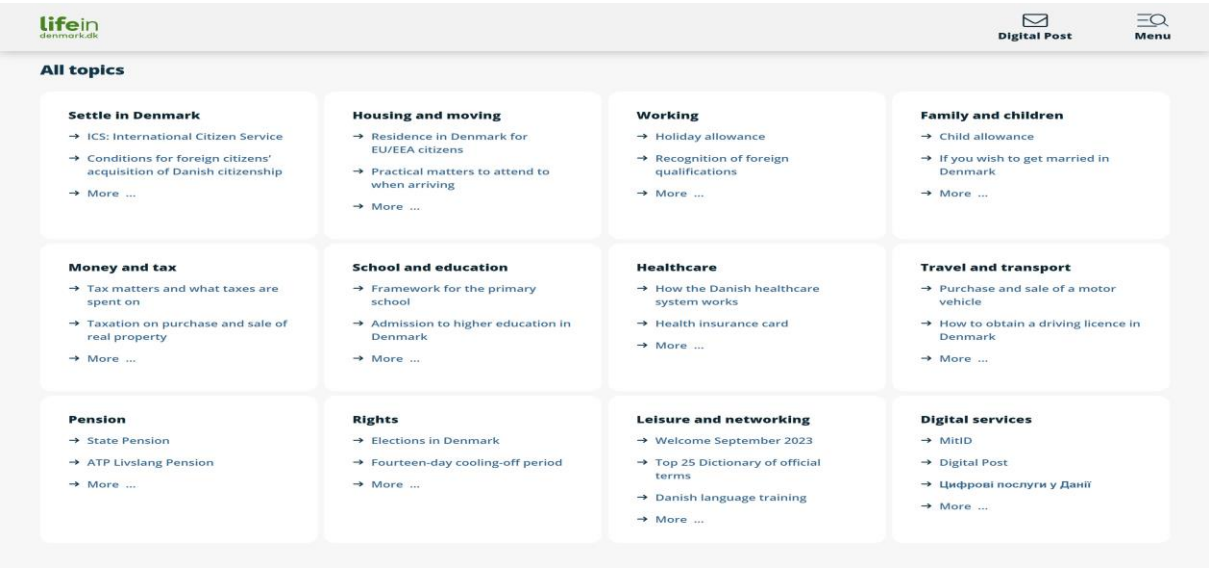

Figure 7: Denmark e-Government web portal "Life in Denmark.dk" [21]

Finland citizen portal "Suomi.fi" is shown in Fig. 8 [22]. In this study we focus on the top two countries (Denmark and Finland) and their citizen portals. First of all, their impact according to the E-Government Development Index is greater than 0.95. Another country, which aspirations were highly estimated is Republic of Korea, however, we failed to access its English web portal version.

Therefore, we obtained the following e-government web portals evaluation results (Table 4). Here SC_i, $i = \overline{1,5}$ describe taxation, education, health, immigration, and employment services.

Table 2
Extracted data from the Denmark e-Government web portal [21]

Service	Link	Fitness
Taxation	More about Money and tax	True
	Tax matters and what taxes are spent on	True
	Taxation on purchase and sale of real property	True
Education	More about Family and children	True
	More about School and education	True
	Child allowance	False
	Framework for the primary school	True
	Admission to higher education in Denmark	True
	Danish language training	True
Health	More about Healthcare	True
	Health insurance card	True
	How the Danish healthcare system works?	True
Immigration	More about Travel and transport	True
	ICS: International Citizen Service	True
	Conditions for foreign citizens	True
	Residence in Denmark for EU/EEA citizens	True
Employment	More about Working	True
	More about Leisure and networking	True
	Framework for the primary school	False
	How the Danish healthcare system works	False

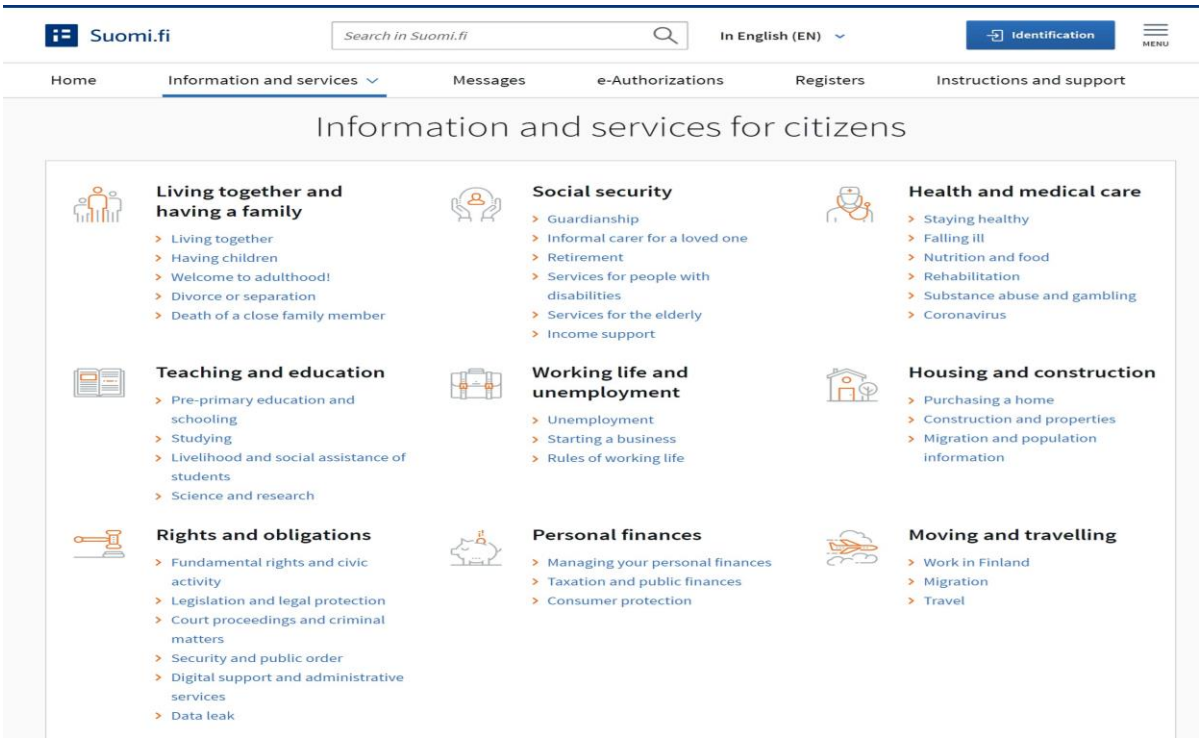

Figure 8: Denmark e-Government web portal "Suomi.fi" [22]

3.2. Citizen Web Portal Data Analysis

Fig. 9 demonstrates the structure of JSON documents produced for Power BI visualization, which contain calculated metrics for assessed citizen web portals.

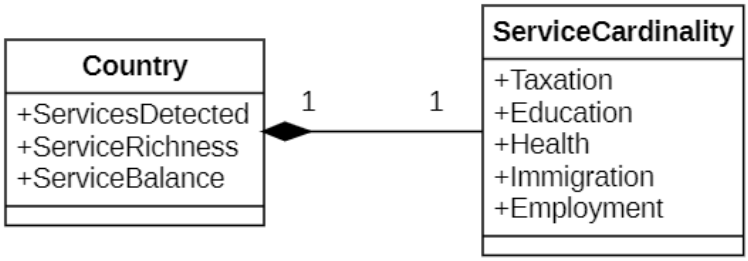

Figure 9: Structure of produced JSON documents

Table 3
Extracted data from the Finland e-Government web portal [22]

Service	Link	Fitness
Taxation	Income support	True
	Managing your personal finances	True
	Taxation and public finances	True
Education	Having children	False
	Pre-primary education and schooling	True
	Studying	True
	Livelihood and social assistance of students	True
Health	Informal career for a loved one	False
	Staying healthy	True
Immigration	For citizens	True
	Travel	True
Employment	Unemployment	True
	Starting a business	True
	Rules of working life	True
	Work in Finland	True
	Contact business advice	True

Table 4
E-government web portals evaluation results

Web portal	SD	SR	SC_1	SC_2	SC_3	SC_4	SC_5	SB
Life in Denmark.dk	5	1.00	0.50	1.00	0.50	0.67	0.67	0.67
Suomi.fi	5	1.00	0.60	0.80	0.40	0.40	1.00	0.64

Here in Fig. 9 we have the following JSON properties:

- "ServicesDetected" represents SD;
- "ServiceRichness" represents SR;
- "ServiceCardinality" represents SC_i, $i = \overline{1,5}$ accroding to IAFEG citizen services of taxation, education, health, immigration, and employment;
- "ServiceBalance" represents SB.

Fig. 10 illustrates the Power BI dashboard developed to visualize JSON-based data and display citizen portal web services assessment results. Analyzing obtained results (Table 4 and Fig. 9 – 10), we can assume that:

- both evaluated "Life in Denmark.dk" and "Suomi.fi" citizen service web portals demonstrate the highest service richness values (1.00), which signalize their general correspondence to 5 citizen services defined by IAFEG [20];
- evaluated citizen web portals focus differently on provided services: "Life in Denmark.dk" is mostly focused on education (1.00), immigration (0.67), and employment (0.67), while "Suomi.fi" on employment (1.00), education (0.80), and taxation (0.60);
- both evaluated citizen web portals have moderate "service balance" scores of 0.67 for "Life in Denmark.dk" and 0.64 for "Suomi.fi", which confirms the previous observations.

Figure 10: Power BI dashboard

Moreover, let us estimate the correlation value between SB and E-Government Development Index (EGDI) values [23]. The obtained Pearson's correlation coefficient [24] value is 1.00, which signalize absolute positive relation between EGDI estimated by UN and "service balance" scores. Finally, let us estimate the accuracy of the proposed information technology using the following formula:

$$Accuracy = \frac{(\sum Fitness = True)}{(\sum Fitness = True) + (\sum Fitness = False)}. \quad (13)$$

Here $\sum Fitness = True$ is the number hyperlinks estimated as correctly categorized against IAFEG services [20], and $\sum Fitness = False$ is vice versa (see Table 2 – 3). Hence, the accuracy of proposed information technology for e-government web portal assessment is 0.80 for Denmark and 0.88 for Finland. However, the total accuracy for both estimated citizen web portals "Life in Denmark.dk" and "Suomi.fi" is 0.83.

Therefore, the proposed information technology allows to obtain accurate (of 83%) e-government web portal assessment results and can be suggested scholars in social political science fields.

4. Conclusion and Future Work

In this paper we proposed the information technology for e-government web portal assessment based on web data extraction techniques. The study aims to improve the processes of e-government web portal assessment by using web harvesting and data analysis approaches. Therefore, we developed algorithms to extract and assess e-government web portals using the proposed E-Government Web Portal Services reference model and evaluation metrics. The software implementation of the proposed technology is based on Python programming language and Power BI data visualization tool. Such a tool allows non-technical users, i.e. social or political science scholars, to configure the desired references models and automatically assess e-government web-portals as part of their studies with the accuracy of 83%.

The following conclusions can be made after the obtained results analysis:

- this approach has a room to identify the differences between e-government web portals and the services they provide;
- consequently, for researchers who study and compare the state of information society formation in different countries, in terms of digital services provision, conducting such experiments can be a useful complement to other data-driven methods;
- the need for interdisciplinary cooperation between social and computer science is increasing and such interdisciplinary studies can benefit both domains with new methods and solutions.

In the future we plan to elaborate metrics proposed to evaluate e-government web portals, as well as conduct a large-scale study, with more government portals and more careful sampling, to identify and study their differences. From the information technology viewpoint, such experiments require advanced techniques to be applied, such as data warehousing, data mining, and data visualization.

5. References

[1] A. Edelmann, T. Wolff, D. Montagne, C. A. Bail, Computational Social Science and Sociology, Annual Review of Sociology 46(1) (2020) 61–81. doi:10.1146/annurev-soc-121919-054621.

[2] D. Lazer et al., SOCIAL SCIENCE: Computational Social Science, Science 323(5915) (2009) 721–723. doi:10.1126/science.1167742.

[3] C. F. Voinea, Political Attitudes: Computational and Simulation Modelling, John Wiley & Sons, 2016. doi:10.1002/9781118833162.

[4] O. Chornenkyi, Use of information and communication technologies for political science research, The Journal of V.N. Karazin Kharkiv National University: Issues of Political Science 42 (2022) 38–44. doi:10.26565/2220-8089-2022-42-06.

[5] B. A. Ojokoh et al., Big data, analytics and artificial intelligence for sustainability, Scientific African 9 (2020) e00551. doi:10.1016/j.sciaf.2020.e00551.

[6] A. Luscombe, K. Dick, K. Walby, Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences, Quality & Quantity 56 (2022) 1023–1044. doi:10.1007/s11135-021-01164-0.

[7] V. Lowndes, D. Marsh, G. Stoker, Theory and Methods in Political Science, 4th ed. Bloomsbury Academic, 2017.

[8] K. L. Anglin, Gather-Narrow-Extract: A Framework for Studying Local Policy Variation Using Web-Scraping and Natural Language Processing, Journal of Research on Educational Effectiveness 12(4) (2019) 685–706. doi:10.1080/19345747.2019.1654576.

[9] J. Osis, U. Donins, Unified Modeling Language: A Standard for Designing a Software, in: Topological UML Modeling, 2017, pp. 3–51. doi:10.1016/B978-0-12-805476-5.00001-0.

[10] R. Tabarés, HTML5 and the evolution of HTML; tracing the origins of digital platforms, Technology in Society 65 (2021) 101529. doi:10.1016/j.techsoc.2021.101529.

[11] B. Gobin-Rahimbux, Evaluation Metrics for Ontology Modules, 2022 IEEE International Conference on Data Science and Information System (ICDSIS), Hassan, India, 2022, pp. 1–6. doi:10.1109/ICDSIS55133.2022.9915950.

[12] A. Boicea, C.-O. Truică, F. Rădulescu, E.-C. Buşe, Sampling strategies for extracting information from large data sets, Data & Knowledge Engineering 115 (2018) 1–15. doi:10.1016/j.datak.2018.01.002.

[13] V. L. Lysytskyi, Y. Y. Morhun, Development of software for effective enterprise product policy creation, Bulletin of National Technical University KhPI Series System Analysis Control and Information Technologies 21 (2018) 59–64. doi:10.20998/2079-0023.2018.21.11.

[14] urllib – URL handling modules. URL: https://docs.python.org/3/library/urllib.html.

[15] re – Regular expression operations. URL: https://docs.python.org/3/library/re.html.

[16] json – JSON encoder and decoder. URL: https://docs.python.org/3/library/json.html.

[17] L. Richardson, beautifulsoup4, PyPI, 2019. URL: https://pypi.org/project/beautifulsoup4/.

[18] H. Zhang, W. Liu, H. Xiong, X. Dong, Analyzing data flow diagrams by combination of formal methods and visualization techniques, Journal of Visual Languages & Computing 48 (2018) 41–51. doi:10.1016/j.jvlc.2018.08.001.

[19] Microsoft Power BI – Interactive Data Visualization BI Tools. URL: https://powerbi.microsoft.com/en-us/.

[20] A. M. Luvembe, H. Mutai, Big Data Framework for Kenya's County Governments, Journal of Computer and Communications 7(1) (2019) 1–9. doi:10.4236/jcc.2019.71001.

[21] The official guide to Life in Denmark. URL: https://lifeindenmark.borger.dk/.

[22] Citizens – Suomi.fi. URL: https://www.suomi.fi/citizen/.

[23] E-Government Development Index (EGDI). URL: https://publicadministration.un.org/egovkb/en-us/About/Overview/-E-Government-Development-Index.

[24] D. L. Hahs-Vaughn, Foundational methods: descriptive statistics: bivariate and multivariate data (correlations, associations), in: International Encyclopedia of Education (Fourth Edition), Elsevier Science, 2023, pp. 734-750. doi:10.1016/B978-0-12-818630-5.10084-3.