CAPAIAN SEMANTIK BERPEMBERAT DOKUMEN SEJARAH BERDASARKAN PERISTIWA

FATIHAH BINTI RAMLI

UNIVERSITI KEBANGSAAN MALAYSIA

CAPAIAN SEMANTIK BERPEMBERAT DOKUMEN SEJARAH BERDASARKAN PERISTIWA

FATIHAH BINTI RAMLI

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEHI
IJAZAH DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

27 April 2018

FATIHAH BINTI RAMLI
P62473

PENGHARGAAN

Bersyukur ke hadrat Illahi kerana dengan izin kurnianya diberi kesabaran dakesihatan yang baik bagi menyiapkan penyelidikan dan penulisan tesis PhD ini. Penghargaan ini ditujukan khas buat semua insan yang terlibat dalam menjayakan perjalanan penyelidikan ini.

Setinggi-tinggi penghargaan dan terima kasih dirakamkan kepada penyelia utama saya, Prof. Dr. Shahrul Azman Mohd Noah atas segala nasihat, dorongan, bantuan dan keprihatinan semasa menyempurnakan tesis ini. Bimbingan, panduan dan tunjuk ajar yang dihulurkan oleh beliau banyak membantu kepada kejayaan tesis ini. Semangat kesabaran serta maklumbalas yang meyakinkan daripada beliau banyak membantu menyempurnakan tesis ini. Ribuan terima kasih diucapkan kepada panel pakar iaitu Prof. Kamarulzaman Yusof yang kini bertugas sebagai pensyarah di Universiti Teknologi Malaysia dan pelajar Institut Pendidikan Guru Kampus Ipoh, saudara Mohamad Nazren Firdaus Bin Mohd Razale dan saudari Siti Adawiyah Binti Mustaffa atas kesudian memberi pandangan dan saranan bagi memurnikan dan menentusahkan beberapa komponen penting mengenai domain sejarah dalam kajian ini.

Pada kesempatan ini, saya juga ingin merakam ribuan terima kasih kepada semua kakitangan Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia dengan segala bantuan dan kemudahan yang diberikan. Saya juga ingin merakamkan terima kasih kepada rakan seperjuangan, terutamanya Tri Basuki Kurniawan yang telah banyak berkongsi ilmu pengetahuan dalam menyempurnakan penyelidikan ini. Segala bantuan serta sokongan beliau sangat dihargai.

Saya juga ingin mengucapkan jutaan terima kasih kepada suami tercinta, Mohd Khairil Abd Rahman di atas segala pengorbanan, dorongan dan kesabaran dalam mengharungi cabaran serta dugaan sepanjang tempoh pengajian ini. Tidak lupa juga kepada ketiga-tiga anak kesayangan saya Muhammad Amirul Hadif, Nur Amni Dhamirah dan Muhammad Aazad Hafiy yang sabar berjauhan daripada saya bagi menamatkan pengajian ini. Penghargaan ini juga ditujukan khas buat kedua-dua ibu bapa dan ahli keluarga saya yang banyak memberi sokongan dan dorongan sepanjang tempoh pengajian ini. Keluarga saya adalah sumber inspirasi bagi menyiapkan tesis ini. Sesungguhnya jasa semua insan yang terlibat dalam menjayakan tesis ini hanya Allah SWT sahaja yang dapat membalasnya. Terima kasih semua.

ABSTRAK

Sistem capaian maklumat generik lazimnya mempunyai kekangan dalam mencapai maklumat bagi suatu domain kompleks seperti biologi, perubatan dan sejarah. Dalam domain sejarah yang menjadi fokus dalam kajian ini sebagai contoh memerlukan perwakilan peristiwa, masa, selang masa, lokasi dan individu bagi memenuhi keperluan maklumat pengguna. Domain sejarah melibatkan rentetan peristiwa yang banyak dan teknik perwakilan peristiwa yang sesuai penting untuk membantu pengguna mencapai maklumat sejarah dengan lebih tepat. Pengguna memerlukan maklumat yang ditakrifkan dengan maknanya dan bukan dengan rentetan teks dan himpunan teks semata-mata. Oleh itu, pembangunan metadata dilihat sebagai satu langkah penyelesaian yang berkesan bagi memudahkan pengaksesan, penyimpanan, dan capaian kepada koleksi dokumen sejarah. Langkah ini turut disokong oleh pendekatan ontologi bagi menerangkan dokumen secara semantik dan membantu meluaskan gelintaran maklumat. Selain itu, ontologi turut dilihat berpotensi untuk mengatasi kekangan semasa sistem capaian maklumat konvensional. Pembinaan model capaian maklumat ini tertumpu kepada capaian dokumen berdasarkan peristiwa sejarah yang mengaplikasikan pendekatan ontologi. Justeru itu, objektif utama penyelidikan ini ialah menyokong capaian dokumen semantik berdasarkan peristiwa dalam domain sejarah dengan menggunakan ontologi. Idea utama penyelidikan ini ialah membangunkan kaedah capaian semantik. Satu senarai indeks dan anotasi dokumen dibina bagi menyokong capaian semantik. Model klasik ruang vektor yang juga memuatkan algoritma pemangkatan diadaptasikan kepada perwakilan berasaskan ontologi bagi menguruskan jumlah maklumat berskala besar. Oleh sebab itu juga, carian dan capaian dokumen ini hanya menumpukan kepada capaian dokumen semantik berdasarkan peristiwa sejarah. Domain ontologi ini dikembangkan melalui kaedah guna semula ontologi sedia ada iaitu ontologi *Simple News and Press Ontologies*(SNaP). Kajian ini memilih beberapa konsep utama dari ontologi seperti ontologi *Event* dan ontologi *Stuff* dari ontologi SNaP bagi membina ontologi yang baru. Konsep yang berpadanan dengan domain spesifik dikekalkan dan seterusnya dikembangkan. Data ujian Battle and Operation of Vietnam War yang mengandungi 134 dokumen digunakan dalam fasa pengujian dan penilaian. Dua puluh kueri dipilih dan diterjemahkan kepada bahasa *Simple Protocol dan RDF Query Language*(SPARQL). Hasil penilaian menunjukkan nilai purata kejituan pendekatan ontologi berasaskan capaian maklumat meningkat sebanyak 34% jika dibandingkan dengan pendekatan berasaskan kata kunci.

ONTOLOGY APPROACH TO RETRIEVING HISTORICAL DOCUMENT BASED ON EVENT

ABSTRACT

Generic information retrieval systems frequently have constraints in retrieving relevant documents involving complex domains such as biology, medical and history. In the history domain for example, which is the focus of this research, required the documents to be represented in terms of event, time, time frame, location and individual in order to fulfil users' information requirements. The history domain involved series of events which need to be represented accurately in order to assists users in retrieving precisely the historical information. Users require information that is defined by its meaning and not merely with string of text and a collection of text alone. Therefore, the development of metadata is one of the approaches to solve the problem for better access, storing and retrieval on the historical documents collection. This development is also supported by the ontology approach to describe the semantics of the document and help broaden the search information. In addition, the ontology approach also has the potential to solve the current constraints of conventional IR system. The construction of IR model is focused on document retrieval based on historical events that apply ontology approach. Hence, the main aim of this research is to support semantic document retrieval in the historical domain based on events by exploiting ontology approach. The main idea is to develop a method to support semantic retrieval. Semantic indexing technique is proposed, and document annotation is created for this purpose. An adaptation of the classic vector-space model for an ontology-based representation is proposed, upon which a ranking algorithm is defined to manage a large scale information sources. Therefore, the search and retrieval of these documents only focus on semantic document retrieval based on historical events. The domain ontology is expanded through reuse of existing ontologies namely Simple News and Press Ontologies(SNaP) ontology. This research selected some ontologies such as Event and Stuff ontologies from SNaP ontology to build a new ontology. The concepts that match to a specific domain is maintained and further developed. The Battle and Operation of Vietnam War test collections consisting of 134 documents are used during testing and evaluation. Twenty textual queries have been chosen which are then transformed into Simple Protocol and RDF Query Language(SPARQL). The result from the evaluation shows the average precision value for the ontology-based IR increased by 34% compared to the conventional approach based on keyword.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI ILUSTRASI

BAB I

PENGENALAN

1.1 PENDAHULUAN

Jumlah maklumat sedia ada dalam Web atau *World Wide Web* terus bertambah dengan pesatnya. Walaubagaimanapun, untuk mendapatkan maklumat yang bertepatan dengan keperluan pengguna agak sukar. Salah satu cara yang berkesan untuk mendapatkan maklumat tersebut adalah melalui kecanggihan enjin gelintaran seperti *Google* dan *Yahoo*. Namun, ciri maklumat yang tidak berstruktur dan jumlah maklumat yang kian meningkat di laman web menyebabkan pengguna menghadapi masalah terutamanya dalam mendapatkan maklumat yang bersesuaian dengan keperluan maklumat pengguna. Enjin gelintaran merupakan aplikasi paling praktikal dalam capaian maklumat. Capaian maklumat merujuk kepada disiplin yang berkaitan dengan mendapatkan maklumat yang sepadan dengan keperluan pengguna. Penyelidikan dalam bidang capaian maklumat berkembang pesat pada tahun 1940-an dengan pembangunan satu kaedah tradisional yang mana melibatkan kaedah perwakilan teks bagi mendapatkan data (Bush 1945; Jones 1997). Kaedah tradisional ini dikenali sebagai capaian teks.

Sejak akhir-akhir ini, dokumen sejarah menjadi fokus utama kepada para sarjana seperti ahli filologi, sejarah, falsafah dan bahasa(Pilz et al. 2006). Kajian mengenai capaian maklumat dokumen sejarah menunjukkan bahawa kebanyakan penyelidikan sedia ada menekankan aspek geografi dan temporal, sistem soal-jawab, dan pengecaman variasi ejaan diantara perkataan moden dan tradisional (Corda 2007; Elena et al. 2010; Gotscharek et al. 2011; Hauser et al. 2007; Jones et al. 2001; Koolen et al. 2006; Mirzaee 2004; Mirzaee et al. 2005; Pilz et al. 2006; Schockaert et al. 2010). Walaubagaimanapun, kebanyakan hasil penemuan penyelidikan ini

menunjukkan keputusan yang kurang memuaskan. Ini adalah kerana kaedah piawai capaian maklumat tidak dapat menyelesaikan sesuatu masalah dengan sendirinya (Hauser et al. 2007). Kaedah piawai ini memerlukan kajian lanjutan untuk mendapatkan keputusan yang lebih baik. Sehubungan dengan itu, kajian capaian maklumat dokumen sejarah masih terbuka luas untuk diterokai.

Dalam kajian ini, pendekatan ontologi diperkenalkan bagi meningkatkan prestasi capaian maklumat dokumen sejarah berdasarkan peristiwa. Dokumen sejarah didefinisikan sebagai peristiwa dan topik yang berkaitan untuk tempoh masa tertentu (Cabo&Llavori 1998). Peristiwa merupakan antara elemen penting yang menjadi tumpuan dalam bidang sejarah (Joseph&Janda 2008). Oleh itu, kajian ini menggariskan bahawa peristiwa sesuai dalam mendapatkan maklumat bagi sesuatu dokumen sejarah secara menyeluruh dan pantas. Capaian dokumen berdasarkan peristiwa merupakan satu proses pertanyaan pengguna dalam bentuk kata kunci dan seterusnya menghasilkan maklumat dalam bentuk fakta bagi sesuatu peristiwa. Dalam menangani masalah penyelidikan ini, ontologi adalah satu model teknologi semantik yang boleh dipadankan dengan dokumen sejarah. Pendekatan ontologi ini menyokong penghasilkan semantik berorientasikan domain yang jelas dari segi pendefinisian konsep, hubungan antara konsep dan tika bagi konsep (Davies et al. 2006). Pendekatan ini membantu proses anotasi semantik, semakan lewa dan capaian dokumen. Ontologi juga dirujuk sebagai graf semantik yang mana perwakilan maklumat dalam bentuk konsep dan hubungan tersusun yang bermakna dan boleh dibaca manusia. Oleh yang demikian, ontologi ini dilihat sebagai salah satu pendekatan yang berkesan bagi mengurangkan jurang semantik.

1.2 LATAR BELAKANG KAJIAN

Capaian maklumat ditakrifkan sebagai menyimpan, menyusun, mewakili, mencari dan mendapatkan maklumat yang sepadan dengan permintaan pengguna (Korfhage 1997). Keberkesanan sistem capaian maklumat antaranya bergantung kepada kaedah perwakilan teks dan domain dokumen yang khusus seperti sejarah yang lazimnya melibatkan keperluan maklumat yang lebih spesifik. Sebagai contoh, ahli sejarah menggunakan pengetahuan, pengalaman dan intuisi untuk menentukan maklumat

yang diperlukan. Ahli sejarah dikehendaki untuk mencari dan mempelajari serta cuba mendapatkan sumber yang berkaitan dengan maklumat yang diperlukan (Elena et al. 2010). Bukti yang ada jelas menunjukkan bahawa ahli sejarah memerlukan repositori sumber sejarah dan alat pembangunan untuk mengakses maklumat yang diperlukan dengan cepat dan teliti (Elena et al. 2010). Oleh itu, capaian maklumat bagi dokumen sejarah merupakan satu bidang penyelidikan penting untuk dikaji.

Dokumen sejarah mengandungi peristiwa dan topik yang berkaitan dalam tempoh masa tertentu (Cabo&Llavori 1998; Llidó et al. 2001). Peristiwa adalah elemen sedia ada dalam dokumen sejarah yang diberi penekanan dalam kajian ini. Kajian ini beranggapan bahawa elemen peristiwa penting dalam sejarah dan ia mempunyai potensi dalam membantu capaian maklumat (Joseph&Janda 2008). Istilah "peristiwa" mempunyai pelbagai maksud. Antaranya ialah peristiwa ditakrifkan sebagai fenomena yang telah berlaku di masa lalu (Shaw et al. 2009). Selain itu dalam kajian sejarah, istilah peristiwa ini didefinisikan sebagai kejadian yang berlaku atau kejadian yang diberikan perhatian (Bedrosian 2008).

Kajian oleh Koolen et al. (2006) berkaitan dengan dokumen lama yang bersejarah menunjukkan bahawa kebanyakan dokumen lama menggunakan istilah atau kata yang tidak lagi digunakan dalam versi moden sejarah. Sehubungan dengan itu, carian dokumen lama ini dengan menggunakan kaedah kata kunci oleh mereka yang bukan pakar lazimnya memberikan hasil yang kurang memuaskan. Ini disebabkan bukan pakar lebih gemar menggunakan kata kunci moden dalam carian dokumen. Koolen et al (2006) mencadangkan pendekatan silang bahasa untuk mendapatkan semula dokumen bersejarah yang bertujuan bagi menyelesaikan kueri pengguna yang menggunakan kata kunci moden.

Dalam pada itu, Pilz et al. (2006) juga mengkaji mengenai warisan budaya dengan menggunakan enjin gelintaran kabur berasaskan peraturan bagi membolehkan pengguna untuk mencapai data teks secara berasingan daripada sistem ortografi. Ortografi ialah huruf dan ejaan yang mementingkan perwakilan pertuturan secara bertulis (Venezky 1999). Petua yang digunakan dalam kajian ini dihasilkan dari analisis statistik, penerbitan sejarah, prinsip linguistik dan pakar pengetahuan. Dalam

kajian ini, penambahbaikan keberkesanan enjin gelintaran dalam kualiti capaian dan fungsian banyak bergantung kepada infleksi dan variasi kata kunci. Penghasilan variasi kata kunci bergantung kepada penggunaan peraturan yang dipilih. Penambahbaikan kajian ini telah memperkenalkan penghasilan petua automatik dan klasifikasi keputusan yang lebih tepat dengan menggunakan kaedah umum persamaan Levenshtein.

Seterusnya, Hauser et al. (2007) pula sangat prihatin tentang capaian maklumat dalam skop warisan budaya untuk dokumen tersembunyi. Dokumen tersembunyi dirujuk kepada buku sejarah dan dokumen. Dalam usaha untuk memelihara dokumen tersembunyi, ia perlu didigitalkan. Manakala, capaian maklumat digunakan sebagai teknik moden untuk mengakses maklumat tersebut. Walau bagaimanapun, teknik pengindeksan piawai dalam capaian maklumat gagal memberikan hasil yang memuaskan untuk sejumlah besar varian ejaan perkataan yang sama (Hauser et al. 2007; Pilz et al. 2006). Hauser et al. (2007) menggunakan kamus khusus, pemadanan generatif berasaskan peraturan dan pemadanan berdasarkan kata persamaan untuk menyelesaikan masalah ini.

Gotscharek et al. (2011) juga menggunakan teknik capaian maklumat untuk mendapatkan dokumen sejarah melalui pencarian di web. Walau bagaimanapun, teknik capaian maklumat piawai yang digunakan gagal menghasilkan keputusan yang memuaskan pada dokumen sejarah kerana kelainan ejaan. Dalam usaha untuk menyelesaikan masalah ini, kata moden yang digunakan dalam pertanyaan ini harus dikaitkan dengan ejaan teks lama dalam dokumen sejarah. Kajian ini telah membina leksikon untuk ejaan teks lama dan disokong oleh prosedur pemadanan. Hasil kajian ini dinilai dari segi ketepatan dan kejituan gabungan kedua pendekatan ini.

Schockaert et al. (2010) menyatakan bahawa capaian maklumat mengenai peristiwa sejarah boleh dicapai menggunakan kueri kekangan temporal. Walaubagaimanapun, perlaksanaan maklumat temporal berstruktur ini sukar dilaksanakan dalam sistem capaian maklumat. Oleh yang demikian, kajian ini mencadangkan satu rangka kerja yang berasaskan pengaburan menggunakan Aljabar selang Allen untuk mengatasi masalah ini. Dengan menggunakan teknik heuristik juga

maklumat temporal dapat diekstrak dari dokumen web. Hasil kajian ini lebih menumpukan kepada keberkesanan kejituan suatu aplikasi pengaburan pentaakulan temporal terhadap kebolehpercayaan pra-pemprosesan maklumat dan isu yang berkaitan dengan kekaburan peristiwa.

1.3 PERMASALAHAN KAJIAN

Latar belakang kajian di atas menunjukkan bahawa terdapat beberapa potensi teknik capaian maklumat boleh diaplikasikan dalam usaha penyimpanan dan capaian dokumen sejarah. Kebanyakan kajian sebelum ini lebih memfokuskan kepada aspek capaian maklumat mengenai kelainan ejaan seperti yang digambarkan dalam Koolen et al. (2006), Pilz et al. (2006), Hauser et al. (2007) dan Gotscharek et al. (2011). Selain itu, terdapat beberapa penyelidikan yang memfokuskan kepada pendigitalan dokumen bersejarah iaitu dokumen tersembunyi sepertimana yang telah dikaji oleh Hauser et al. (2007) dan Järvelin et al. (2016). Namun, capaian maklumat berdasarkan peristiwa kurang mendapat perhatian walaupun elemen peristiwa adalah penting dalam dokumen sejarah kerana fakta menunjukkan bahawa kebanyakan dokumen sejarah mengandungi peristiwa selain daripada topik. Tambahan pula, ahli sejarah biasanya tertarik dengan peristiwa sesuatu topik dan mereka berharap agar pertanyaan disusun secara spesikfik mengikut peristiwa (Grossner 2010; Shaw 2010). Namun, sistem konvensional capaian maklumat tidak dapat menyokong keperluan pembangunan web semantik disebabkan hanya menggunakan perwakilan dokumen *bag-of-words* yang ringkas. Model capaian maklumat *bag-of-words* mewakili dokumen dengan hanya satu struktur berdasarkan set perkataan dan tidak membenarkan permodelan hubungan antara subset kata(Liu et al. 2007). Gabungan pendekatan *bag-of-words* dan TF-IDF dilihat boleh membantu menangkap semantik sesuatu dokumen namun apabila melibatkan repositori yang besar gabungan kedua pendekatan ini memberi keputusan yang kurang memuaskan kerana semantik antara kata dalam dokumen akan hilang. Oleh yang demikian, salah satu cara yang dapat membantu menyokong keperluan tersebut ialah dengan menggunakan pendekatan ontologi yang berfungsi meningkatkan perwakilan dokumen secara semantik bagi capaian maklumat yang lebih tepat(Dragoni et al. 2012). Salah satu alternatif adalah dengan menggunakan pendekatan capaian maklumat berasaskan ontologi.

Penggunaan ontologi telah diajukan sebagai salah satu motivasi untuk web semantik yang mampu menyokong carian semantik bagi repositori dokumen dan hasil kajian menunjukkan peningkatan yang jelas dalam capaian maklumat berbanding enjin gelintaran berdasarkan kata kunci(Castells et al. 2007). Contohnya, Mirzaee et al. (2005) dan Mirzaee (2004) membangunkan model ontologi formal untuk menentukan konsep dan hubungan bagi dokumen sejarah. Namun, kajian model ontologi sejarah ini mengekstrak konsep dan ciri yang penting secara manual dari satu buku sejarah sahaja. Kajian ini hanya menunjukkan bagaimana konsep ontologi boleh digunakan dalam menganotasi dokumen untuk menyokong sistem soal-jawab (Mirzaee 2004).

Selain itu, Corda (2007) menggunakan ontologi sebagai teknologi pintar yang boleh memperkayakan akses kepada pelbagai sumber digital secara semantik. Corda (2007) membangunkan rangka kerja untuk menentukan konsep dan taakulan tentang Sejarah Sains dengan menggabungkan teori Davidson mengenai peristiwa dan logik selang Allen. Walau bagaimanapun, kajian ini hanya mengendalikan satu kajian kes spesifik dalam Sejarah Sains yang menggambarkan bagaimana untuk menentukan konsep dan taakulan tentang domain sejarah. Kajian ini telah memberikan perhatian khusus tentang rangka kerja mengenai bagaimana ontologi boleh digunakan dalam pemodelan masa dan hubungan temporal dalam Sejarah Sains untuk menyokong sistem soal-jawab.

Walaupun kerja pemodelan ontologi sejarah telah dilaksanakan dalam beberapa kajian seperti Mirzaee (2004), Mirzaee et al. (2005) dan Corda (2007), namun kajian tersebut kurang memberi perhatian kepada kepentingan bidang capaian maklumat. Kebanyakan kerja penyelidikan hanya ditumpukan dalam membangunkan ontologi sejarah bagi menjawab soalan kecekapan tertentu. Oleh yang demikian, kebanyakan kajian di atas mendapati bahawa capaian maklumat berdasarkan peristiwa yang disokong oleh ontologi dapat memperkaya capaian dokumen sejarah. (Gotscharek et al. 2011; Hauser et al. 2007; Koolen et al. 2006; Pilz et al. 2006).

1.4 PERSOALAN KAJIAN

Persoalan kajian banyak menitikberatkan komponen berkaitan model ontologi bagi capaian dokumen sejarah berdasarkan peristiwa:

1) Apakah model ontologi yang sesuai untuk domain sejarah dalam membantu pengguna mendapatkan maklumat dengan lebih tepat?

2) Apakah kaedah yang sesuai untuk mengekstrak dan mempopulasikan model ontologi ini?

3) Bagaimana pra-pemprosesan semi-automatik terhadap anotasi teks dilaksanakan?

4) Bagaimana elemen semantik dibangunkan dari kandungan teks tanpa struktur?

5) Bagaimana model capaian maklumat berasaskan ontologi memberi kesan kepada prestasi capaian?

1.5 OBJEKTIF KAJIAN

Tujuan utama kajian ini adalah untuk membina model capaian maklumat berdasarkan peristiwa untuk dokumen sejarah berasaskan ontologi. Bagi menyokong matlamat ini, objektif berikut digariskan:

1) Merekabentuk dan membina model ontologi yang boleh menyokong pertanyaan dan capaian dokumen sejarah berasaskan semantik.

2) Merekabentuk dan menambahbaik peraturan bagi mengekstrak dan mempopulasikan ontologi.

3) Menguji dan menilai ketepatan hasil capaian dokumen sejarah berdasarkan peristiwa menggunakan ontologi.

1.6 KEPENTINGAN KAJIAN

Kepentingan kajian ini adalah dalam bidang capaian maklumat dan teknologi semantik. Kajian ini dilaksanakan bagi mengaplikasikan pendekatan ontologi dalam domain sejarah untuk mendapatkan semula dokumen sejarah dengan lebih spesifik dan tepat. Pendekatan ontologi digunakan bagi merapatkan jurang semantik dalam

dapatan semula teks. Seterusnya, pembangunan model ontologi ini memfokuskan carian berdasarkan peristiwa yang berlaku. Model ontologi berdasarkan peristiwa ini dapat menyelesaikan isu bagi mendapatkan kandungan maklumat yang relevan dan menepati kehendak pengguna. Kajian ini memberi manfaat kepada pakar sejarah dan semua pengguna yang menggunakan dokumen sejarah dalam rutin harian mereka. Perlaksanaan model ontologi baru ini dapat memberi impak yang signifikan dalam bidang ini dan seterusnya membantu meningkatkan keberkesanan dalam sistem capaian dokumen.

1.7 SKOP KAJIAN

Skop kajian ini tertumpu kepada capaian dokumen berdasarkan peristiwa yang mengaplikasikan pendekatan ontologi. Dokumen dan domain kajian ini adalah dalam bahasa Inggeris yang mana bersesuaian dengan domain spesifiknya. Domain spesifik kajian ini ialah domain sejarah. Domain sejarah ini hanya memfokuskan kepada Peperangan Vietnam yang berlaku pada Perang Dunia ke-2[1]. Pada abad keenam, peperangan Vietnam ini menjadi tumpuan pertikaian dan titik konflik utama di antara kuasa besar seperti Amerika dan Soviet berikutan Peperangan Dunia Kedua(Lawrence 2010). Ini adalah kerana kuasa besar di seluruh dunia melihat kegawatan politik di Vietnam sebagai perkara yang penting. Sumber maklumat domain ini boleh didapati di laman Wikipedia. Sumber maklumat domain ini turut digunakan dalam satu penyelidikan lepas iaitu kajian Schockaert et al. (2010). Oleh sebab itu, kajian ini yakin sumber maklumat domain ini boleh dipercayai dan dipilih bagi menjalankan kajian ini. Sasaran pengguna kajian ini adalah tertumpu kepada pakar sejarah dan pengguna yang menggunakan dokumen sejarah dalam rutin kerja harian mereka. Pakar sejarah memerlukan repositori sumber sejarah bagi memudahkan mereka mendapatkan maklumat sejarah dengan tepat dan cepat. Oleh itu, kajian ini memerlukan pemahaman dan penerokaan mengenai domain spesifik dan kaedah pembangunan ontologi dalam capaian maklumat sejarah. Selain itu, kajian ini turut mengkaji ontologi sedia ada untuk mengguna semula ontologi sebagai panduan kepada pembangunan ontologi baru ini di samping membantu memperkayakan istilah

[1] http://en.wikipedia.org/wiki/Category:Battles_and_operations_of_the_Vietnam_War

asas konsep dan hubungan ontologi seperti SNaP[2]. Seterusnya, penerokaan alatan yang berkaitan turut dilakukan bagi membantu aktiviti utama seperti alatan GATE[3] dan TopBraid Composer[4]. Alatan ini dapat membantu aktiviti pengekstrakan dan pembangunan ontologi. Penilaian akhir kajian ini melibatkan penilaian ketepatan capaian dokumen sejarah dengan menggunakan pendekatan ontologi terutamanya dalam elemen peristiwa.

1.8 METODOLOGI KAJIAN

Metodologi kajian ini merangkumi empat komponen utama. Komponen pertama ialah perolehan dan analisis pengetahuan. Komponen ini menjelaskan bagaimana kajian literatur dan analisis dijalankan bagi mengenalpasti masalah dan kelemahan yang ada dalam bidang teknologi semantik dan capaian maklumat. Pengenalpastian masalah dan ringkasan literatur ini banyak merujuk kepada buku, jurnal, artikel dan tesis. Aktiviti analisis ini dijalankan bagi mengkaji model ontologi bagi domain sejarah dan mengenalpasti apakah kelemahan model tersebut serta apakah pendekatan ontologi yang sesuai untuk dibangunkan bagi menangani masalah yang timbul. Kemudian, aktiviti diteruskan dengan mendefinisikan keperluan dan spesifikasi yang diperlukan dalam pembangunan ontologi.

Komponen kedua ialah pangkalan pengetahuan. Ia merangkumi proses rekabentuk dan pembangunan model ontologi. Proses rekabentuk ini mengkaji tentang model ontologi semasa yang digunakan dalam dapatan semula dokumen. Proses ini juga termasuk merekabentuk proses pra-pemprosesan maklumat secara automatik yang mana melibatkan proses pra-pemprosesan konsep penting dalam pembangunan ontologi. Seterusnya, rangka kerja secara menyeluruh dirangka bagi pra-pemprosesan maklumat dan pembangunan model ontologi untuk dapatan semula dokumen. Dengan adanya rangka kerja ini, proses pra-pemprosesan maklumat dan pembangunan ontologi boleh dilaksanakan. Kedua-dua proses ini disokong oleh keperluan sistem, bahasa ontologi dan alat ontologi. Dalam proses ini, ontologi dibangunkan

---

[2] http://data.press.net/ontology/

[3] https://gate.ac.uk/

[4] https://www.topquadrant.com/tools/modeling-topbraid-composer-standard-edition/

menggunakan alat ontologi TopBraid Composer. Kemudian, model ontologi disokong oleh alat pra-pemprosesan maklumat bagi melaksana pra-pemprosesan teks secara automatik. Proses pra-pemprosesan maklumat ini melibatkan pembangunan peraturan tatabahasa mengenai teks. Di akhir komponen ini satu model ontologi terhasil.

Seterusnya ialah komponen ketiga iaitu proses perlaksanaan. Proses perlaksanaan melibatkan pemetaan model ontologi kepada dokumen yang berkaitan. Proses pemetaan ini dilaksanakan dengan mengimport dokumen dari fail lokal terus ke dalam model ontologi. Kemudian, proses pemetaan ini diuji menggunakan bahasa SPARQL bagi memastikan dokumen yang dipadan menepati kehendak pengguna. Output proses perlaksanaan ini merupakan input kepada proses seterusnya iaitu proses penilaian.

Proses penilaian merupakan komponen keempat dalam kajian ini. Proses ini menerima output dari fasa perlaksanaan bagi mendapatkan keputusan ketepatan dan perolehan kembali. Keputusan tersebut dianalisis dan dibandingkan dengan carian berdasarkan kata kunci piawai bagi menentukan ketepatan dan keberkesanan pendekatan yang dicadangkan. Output bagi proses ini diterangkan secara terperinci dalam Bab 6 Penilaian dan Analisis. Di akhir kajian ini, terdapat beberapa cadangan dicadangkan sebagai kerja masa hadapan.

1.9 GAMBARAN KESELURUHAN TESIS

Kajian ini mempunyai enam bab. Bab yang seterusnya adalah seperti dibawah.

Bab 2 ialah Tinjauan Literatur yang membincangkan tentang kajian-kajian lepas yang berkaitan dengan model ontologi dan capaian maklumat. Tinjauan literatur ini memfokuskan kepada isu-isu semasa dalam ontologi domain spesifik dan sejauh mana ontologi penting dalam membantu dapatan semula teks. Selain itu, kajian ini juga membincangkan secara terperinci mengenai domain spesifik dan apakah kepentingan ontologi dan capaian maklumat dalam domain tersebut. Perbincangan ini dianalisis secara terperinci bagi mendapatkan kefahaman tentang jurang utama yang wujud dalam kebanyakan isu-isu yang timbul dalam kajian sebelum ini. Ia juga termasuk analisis tentang model terpilih yang fokus kepada domain spesifik yang

mana telah dibangunkan oleh para penyelidik lain. Kesimpulan dari tinjauan ini membantu menjawab objektif pertama kajian ini yang mana merangkumi tahap pencapaian (*state of the art*) dan mengenalpasti jurang masalah dalam model ontologi bagi domain spesifik.

Bab 3 ialah kaedah kajian yang membincangkan metodologi kajian yang telah digunakan. Bab ini memfokuskan kepada pembangunan ontologi untuk dapatan semula dokumen sejarah berdasarkan peristiwa. Seperti mana yang telah dibincangkan sebelum ini, kaedah kajian mempunyai empat komponen iaitu perolehan pengetahuan dan analisis, merekabentuk pangkalan pengetahuan, perlaksanaan dan penilaian model ontologi.

Bab 4 ialah Model ontologi untuk dapatan semula dokumen sejarah berdasarkan peristiwa. Bab ini menerangkan bagaimana membangunkan satu model ontologi berdasarkan capaian maklumat yang efektif yang mana menggunakan kata kunci peristiwa sejarah bagi mencari dan mendapatkan maklumat secara lebih tepat dan spesifik. Bab ini juga menerangkan model pangkalan pengetahuan untuk sistem semantik dapatan semula teks. Selain itu, proses pembangunan model ontologi peristiwa juga diterangkan secara terperinci yang mana ontologi ini kemudiannya digunakan dalam proses pembangunan kaedah-kaedah baru dalam proses pra-pemprosesan dan populasi ontologi bagi mendapatkan semula dokumen sejarah yang spesifik. Oleh itu, bab ini dapat memberi jawapan kepada objektif kedua kajian ini.

Bab 5 ialah Penilaian dan Analisis yang mana mempersembahkan prestasi model ontologi untuk dapatan semula dokumen sejarah berdasarkan peristiwa. Prestasi model ontologi ini dipersembahkan dari segi keputusan ketepatan dan perolehan kembali maklumat serta F-measure. Model ontologi baru ini dibandingkan dengan sistem carian piawai yang berdasarkan kata kunci dan memaparkan keputusan yang lebih efektif dengan mengintegrasikan ontologi ke dalam capaian maklumat.

Bab 6 ialah Kesimpulan dan Kajian Lanjutan yang mana menyimpulkan tesis ini dengan membangkit isu-isu yang timbul dan penemuan-penemuan baru daripada kajian ini. Selain itu, bab ini juga menerangkan tentang sumbangan yang telah

didapati daripada tesis ini. Di akhir bab ini, beberapa cadangan dibuat sebagai kajian lanjutan kepada pembaca tesis ini sekiranya berminat untuk melanjutkan kajian ini.

BAB II

KAJIAN LITERATUR

2.1 PENGENALAN

Kajian literatur merupakan kajian awal yang dilakukan dalam sesuatu bidang yang ingin dikaji. Pada peringkat pertama, kajian ini meninjau mengenai ciri domain sejarah. Selain itu, kajian ini turut meninjau pendekatan ontologi dan penggunaannya dalam bidang sejarah serta fenomena rekabentuk serta proses pembangunan ontologi. Seterusnya, kajian ini meninjau ontologi sebagai satu cara penyelesaian kepada capaian maklumat. Tinjauan ke atas ontologi dijalankan berkenaan pemodelan ontologi dalam kajian lepas.

Model ontologi sejarah yang dikaji ini adalah model yang memfokuskan kepada elemen peristiwa. Beberapa kajian lepas berkenaan permodelan ontologi sejarah berdasarkan peristiwa dilihat secara khusus. Di samping itu juga, tinjauan terhadap domain sejarah yang merupakan domain spesifik kajian ini turut dikaji berdasarkan kajian lepas berkenaan pemodelan ontologi sejarah berdasarkan peristiwa. Kesemua tinjauan ini dapat memberi maklumat kepada penyelidik berkenaan latar belakang kajian serta membantu menyelesaikan masalah dan kekangan yang timbul dalam kajian lepas dengan memberi nilai tambah pada kajian tersebut.

2.2 CIRI DOMAIN SEJARAH

Sejarah didefinisikan sebagai penemuan, pengumpulan, penyusunan, dan pembentangan maklumat mengenai peristiwa yang lepas (Joseph&Janda 2008). Sejarah boleh dirujuk sebagai tempoh masa selepas tulisan dibuat. Ia adalah satu bidang penyelidikan yang menggunakan naratif untuk memeriksa dan menganalisis

urutan peristiwa, dan kadang-kadang cuba untuk menyiasat secara objektif corak sebab dan akibat yang menentukan sesuatu peristiwa. Di samping itu, sejarah juga dikenali sebagai rekod kronologi peristiwa penting yang menjelaskan sebab-sebab sesuatu peristiwa dan beberapa cabang ilmu yang direkod serta menerangkan peristiwa masa lalu (Merriam-Webster 2012).

Kebanyakan definisi sejarah menerangkan peristiwa sebagai sebahagian daripada sejarah. Sejarah dan peristiwa adalah berkait rapat antara satu sama lain. Peristiwa tidak boleh dipisahkan daripada sejarah dan sejarah tidak boleh diasingkan daripada peristiwa. Justeru, Kumar (2011) menyatakan bahawa sejarah merupakan rekod peristiwa lalu yang mempuyai sebab-sebab tersendiri dan berhubungan diantara satu sama lain. Selain itu, JosephandJanda (2008) turut menjelaskan bahawa kebanyakan kajian sejarah memfokuskan kajian mereka berkenaan peristiwa dan pembangunan yang berlaku pada sesuatu masa tertentu. Istilah peristiwa mempunyai beberapa maksud, yang mana salah satu daripadanya didefinisikan sebagai satu fenomena yang berlaku pada masa yang lepas (Shaw et al. 2009). Seterusnya Kumar (2011) mendefinisikan peristiwa sebagai saling berkaitan. Dalam kajian Ibrahim (1994) turut menjelaskan bahawa banyak peristiwa yang berlaku pada masa lalu berkaitan antara satu sama lain. Dengan itu, peristiwa boleh dikategorikan sebagai salah satu elemen utama dalam domain sejarah. Oleh sebab itu, kajian ini memilih peristiwa sebagai perwakilan data untuk domain sejarah. Peristiwa merupakan sesuatu yang kompleks untuk dimodelkan kerana ia perlu mewakilkan maklumat temporal, ruang, pengalaman, struktur dan aspek sebab-dan-akibat sesuatu peristiwa berlaku (Wang&Zhao 2012). Sehubungan dengan itu perwakilan ontologi merupakan antara teknik yang mampu untuk memodelkan peristiwa dan domain sejarah.

2.3 ONTOLOGI

Istilah ontologi berasal daripada bidang falsafah yang bermaksud penerangan mengenai sifat dan kewujudan serta kategori asas dan hubungan sesuatu benda (Gruber 2008). Sementara itu, istilah ontologi dalam konteks teori pula didefinisikan mengikut istilah kecerdasan buatan (AI) iaitu ‘spesifikasi formal yang jelas (eksplisit) tentang sesuatu pengkonsepan’ (Gruber 1993). Ia menyediakan perbendaharaan kata

yang boleh dikongsi dan digunakan untuk memodelkan domain seperti jenis objek atau konsep yang wujud serta sifat dan hubungan mereka (Arvidsson&Flycht-Eriksson 2008). Definisi ini menerangkan pembentukan konsep dan hubungan untuk domain tertentu yang dihasilkan oleh ontologi. Oleh kerana pada masa kini aplikasi ontologi digunakan secara meluas untuk bidang yang berbeza seperti undang-undang, e-dagang, pengurusan pengetahuan, Web Semantik dan sejarah, maka definisi Gruber telah berkembang. Davies et al. (2006) merumuskan dua perkara penting berdasarkan definisi Gruber. Pertama ialah pengkonsepan merupakan konsep formal yang mana disokong oleh penaakulan komputer. Kedua ialah rekabentuk sesuatu ontologi adalah untuk domain tertentu sahaja.

2.3.1 Ontologi dan Klasifikasi

Definisi ontologi menjelaskan bahawa ia merupakan satu kriteria konkrit yang dapat memenuhi keperluan entiti bagi mendefinisikan pengabstrakan pelbagai aras (Davies et al. 2006). Beberapa klasifikasi mengenai ontologi telah dijelaskan dalam literatur (Borgo 2007; Gómez-Pérez, Corcho, et al. 2004; Lassila&Mcguinness 2001). Kajian ini menggunakan klasifikasi berdasarkan skop sesuatu ontologi atau granulariti domain. Ontologi yang bagus mesti mematuhi beberapa kriteria iaitu:

a. Abstrak

Ontologi dirumuskan secara umum. Ia bersifat umum bagi membolehkan proses guna semula dalam pelbagai bidang domain.

b. Boleh guna

Ontologi boleh digunakan dalam pelbagai konteks semantik. Pengguna ontologi tidak boleh mengubah tika sedia ada sewenangnya kerana ia boleh memberi kesan kepada sesuatu entiti.

c. Boleh ditentusah

Setiap kriteria individu boleh dinilai. Ontologi juga didefinisikan sebagai satu sistem abstrak, boleh guna dan entiti boleh ditentusah. Walaubagaimana pun, ontologi juga mesti memenuhi kriteria-kriteria lain yang mana ia merupakan ciri-ciri sesuatu entiti.

d. Lengkap

Semua konsep dan entiti diekstrak dalam bentuk teks dari dokumen sejarah dan dipadankan kepada situasi tertentu.

e. Unik

Ontologi didefinisikan dengan sempurna. Contohnya, sekiranya satu ontologi digandingkan ke dalam sesuatu sistem, ia akan memberi keputusan yang sama dan tidak menjejaskan instance sedia ada.

f. Tersusun

Semua entiti dalam ontologi disusun secara sistematik.

g. Efisien

Ontologi tidak memerlukan peralatan sokongan. Sesuatu aplikasi boleh dibangunkan mengikut masa yang diberi.

Menurut MadsenandThomsen (2009), definisi ontologi dan klasifikasi adalah berbeza mengikut tujuan penggunaannya dalam struktur pengetahuan. Secara asasnya, ontologi merupakan satu model manakala klasifikasi ialah satu sistem. Peranan model ontologi ialah membantu menyediakan satu perwakilan pengetahuan ringkas tentang sesuatu fenomena. Manakala klasifikasi pula memfokuskan kepada pembahagian sesuatu fenomena kepada pecahan kelas yang berlainan bagi menyokong pembangunan model ontologi. Kesimpulannya, klasifikasi merupakan asas kepada pembangunan sesuatu ontologi.

2.3.2 Pembinaan Ontologi

Secara keseluruhannya, pembinaan ontologi memerlukan kemahiran pakar dan mengambil masa yang lama (Weng et al. 2006). Pembinaan ontologi mementingkan klasifikasi konsep, hubungan dan cara melaksanakan ontologi. Kini, kebanyakan pembinaan ontologi adalah secara semi-automatik iaitu penakrifan konsep dan hubungan dilaksanakan secara hipotesis. Oleh yang demikian, fasa rekabentuk dan penilaian adalah sebahagian daripada keperluan asas pembinaan ontologi.

2.3.3 Pendekatan Rekabentuk Dan Penilaian Ontologi

Kejuruteraan ontologi menitik beratkan rekabentuk asas, pengubahsuaian, aplikasi dan penilaian ontologi. Terdapat beberapa kriteria yang menjadi panduan kepada proses rekabentuk ontologi (G'abor 2007; Gruber 1993) dan ia juga digunakan untuk menilai rekabentuk ontologi:

a. Kejelasan

Kejelasan di sini merujuk kepada pemilihan istilah ontologi yang jelas. Setiap istilah yang dipilih mempunyai hubungkait dengan skop dan objektif pembangunan ontologi. Jumlah tafsiran istilah adalah terhad kepada skop tersebut.

b. Pertautan

Semua cadangan yang terhasil dari definisi ontologi dan aksiom mempunyai pertautan dengan definisi ontologi dan aksiom yang lain. Kesimpulan yang terhasil hendaklah konsisten dengan definisi dan aksiom sedia ada serta jelas dan logik.

c. Boleh dikembangkan

Pengetahuan yang sentiasa berkembang menjadikan ontologi sebagai landasan perkongsian maklumat. Ontologi sedia ada boleh digunapakai dan dikembangkan dalam pelbagai bidang. Pendefinisian istilah dalam setiap ontologi baru hendaklah mudah dilaksanakan berdasarkan perbendaharaan kata sedia ada tanpa perlu mengubah definisi ontologi tersebut.

d. Pengekodan bias yang minimum

Matlamat ontologi ialah untuk mewakili fakta yang benar. Malah ontologi juga bersifat berkompromi mengenai sebarang penambahbaikan terhadap ontologi. Sebagai contoh, pengkonsepan hendaklah dilakukan pada fasa pengetahuan bukan hanya simbolik.

e. Komitmen ontologi yang minimum

Komitmen besar dalam ontologi menjadikan struktur ontologi lebih tegar dan menghadkan jumlah pengguna ontologi. Komitmen ontologi hendaklah minima tetapi cukup untuk menyokong ativiti perkongsian maklumat. Sebagai contoh, sesuatu ontologi itu dapat menyokong komunikasi berkesan secara konsisten berdasarkan pengkonsepan domain dan mudah dikembangkan oleh individu.

Semua kriteria di atas berpotensi sebagai panduan kepada permulaan rekabentuk ontologi atau pengubahsuaian ontologi. Sebagai contoh, kriteria ini digunakan bagi mendapatkan maklumbalas mengenai aplikasi, penilaian terhadap ciri-ciri ontologi ataupun pertukaran domain.

Penggunaan ontologi terus berkembang dalam pelbagai bidang kajian. Antaranya ialah permodelan masalah dan domain dalam bidang perniagaan, sains kesihatan, sukan, berita, sejarah dan sebagainya. Dengan itu, metodologi bagi rekabentuk ontologi dan penilaian adalah sangat penting. Menurut GrüningerandFox (1995), setiap ontologi membenarkan perkongsian terminologi dan beberapa batasan dalam setiap objek yang telibat. Di samping itu, ontologi menyediakan satu koleksi "*easy to re-use*" mengenai pengkelasan objek bagi permodelan masalah dan domain(Gruninger 1996). Oleh yang demikian, satu mekanisma telah dirangka sebagai panduan untuk merekabentuk ontologi dan juga satu rangka kerja dibina bagi menilai ketepatan ontologi.

2.3.4 Bahasa Ontologi

Bahasa ontologi ialah bahasa formal yang digunakan untuk membina ontologi. Bahasa ontologi membenarkan pengekodan pengetahuan sesuatu domain dan lazimnya juga melibatkan petua penaakulan yang menyokong pemprosesan pengetahuan tersebut. Pelbagai bahasa ontologi diperkenalkan seperti: CycL, KL-One, Ontolingua, F-Logic, OCML, LOOM, Telos, RDF(S), OIL, DAML+OIL, XOL, SHOE, dan *Web Ontology Language* (OWL) Walau bagaimanapun OWL yang dihasilkan oleh *World Wide Web Consortium* (W3C) telah dipersetujui sebagai bahasa piawai ontologi. OWL membantu menyediakan mekanisme bagi membentuk semua komponen ontologi seperti konsep, tika, hubungan dan aksiom (Davies et al. 2006).

Ciri OWL terdiri daripada koleksi operator yang mana berfungsi bagi menerangkan konsep seperti operator boolean iaitu persilangan, kesatuan dan pelengkap serta kuantiti yang jelas bagi sifat dan hubungan. Dengan adanya koleksi operator ini, ia mampu menentukan ciri sesuatu sifat seperti ketransitifan atau domain dan julat; penghasilan semantik memudahkan penggunaan inferens dan penaakulan automatik; penggunaan *Universal Resource Identifier* (Fernández-López et al.) bagi menamakan konsep dan ontologi; mekanisma untuk mengimport ontologi luar dan keserasian seni binanya dengan *World Wide Web* khususnya bahasa perwakilan lain seperti RDF dan skema RDF (Bechhofer 2009).

Berdasarkan penerimaan OWL sebagai bahasa piawaian untuk menakrifkan ontologi, kajian ini menggunakan OWL dalam membina dan menakrifkan ontologi domain sejarah yang dicadangkan.

2.3.5 Ontologi Sejarah dan Peristiwa

Kajian ini memilih untuk mengguna semula ontologi sedia ada untuk pembinaan ontologi baru. Beberapa ontologi sedia ada telah dikenalpasti bagi membantu memperbaiki dan mengembangkan domain spesifik. Dalam kajian ini, penggunaan semula ontologi sedia ada penting bagi membina ontologi baru dari awal terutamanya dari segi mengenalpasti konsep yang perlu dalam ontologi baru. Terdapat beberapa ontologi sedia ada yang dikenalpasti seperti berikut:

a. Ontologi STOLE

STOLE adalah ontologi rujukan yang menyediakan kosa kata istilah dan hubungan untuk memodelkan spesifik domain dengan jelas. STOLE ontology menggunakan sejarah Pentadbiran Awam Itali sebagai domain khusus. Matlamat utama ontologi STOLE adalah untuk mempunyai model reka bentuk yang jelas mengenai konsep sejarah dan mencari pandangan mengenai bidang tertentu. STOLE bertujuan untuk mengumpulkan maklumat mengenai jurnal yang paling relevan mengenai sejarah undang-undang pentadbiran awam di Itali yang diterbitkan antara 1848 dan 1946. Pembinaan ontologi STOLE terdiri daripada tiga fasa utama: 1) Pengenalan konsep utama, 2) Pengenalan bahasa yang betul dan Pelaksanaan Tbox, 3) Populasi Ontologi (Adorni et al. 2015). Pada fasa pertama, konsep utama yang terlibat dalam domain tertentu mesti ditakrifkan oleh pakar domain. Pakar domain menyediakan penjelasan semantik manual yang akan ditambah kepada ontologi melalui program JAVA. Seterusnya, mereka mengelaskan semua data yang berkaitan dengan dokumen sejarah dan hasil semua konsep akan dilihat dalam bentuk taksonomi yang terdiri daripada tiga elemen seperti yang ditunjukkan dalam Jadual 2.1 dan menunjukkan saiz ontologi STOLE yang dikira oleh PROTEGE. Akhirnya, populasi ontologi dijalankan untuk membantu mengisi entiti yang hilang secara automatik di Abox dengan penjelasan semantik. Ontologi STOLE boleh diakses oleh orang ramai dan boleh dianggap sebagai ontologi yang boleh diperkembangkan. Ontologi ini dibina dalam bahasa Itali.

Jadual 2.1 Taksonomi Ontologi STOLE

Elemen	Contoh
Data pengarang artikel	Nama, nama keluarga, biografi
Data jurnal dan artikel	Tajuk artikel, nama jurnal, tarikh dan topik dalam artikel
Data mengenai fakta relevan dan orang yang berkaitan dipetik dalam artikel	Orang berkaitan, peristiwa sejarah, institusi

Jadual 2.2 Statistik Tbox Ontologi STOLE

Elemen	Saiz
Kelas	14
Aksiom	440
Sifat Objek	30
Sifat Data	39

b. Ontologi Peristiwa

Hyvönen et al. (2007) menyatakan bahawa portal semantik untuk warisan budaya memerlukan ontologi peristiwa kerana tiga sebab: 1) Peristiwa memerlukan pengidentifikasi ontologi, *Uniform Resource Identifier (URI)* untuk membina koleksi metadata, 2) Peristiwa penting dalam mewujudkan hubungan semantik antara kandungan budaya dan 3) Peristiwa sejarah penting untuk membentuk tulang belakang sejarah kronologi. Hyvönen et al. (2007) mengembangkan ontologi peristiwa menggunakan sejarah Finland sebagai domain khusus. Ontologi peristiwa sejarah adalah berdasarkan garis masa yang dibuat oleh rangkaian Agricola dan dimanfaatkan sebagai sebahagian daripada portal semantik "*CultureSampo-Finnish Culture on the Semantic Web*", sebuah sistem susulan lintas-domain mengenai Muzium Finland. Pengelasan peristiwa didasarkan pada garis masa temporal dan dimensi lain seperti jenis peristiwa iaitu perang, upacara kemahkotaan atau cabang sejarah iaitu sejarah politik, sejarah sains. Mereka menganotasi secara manual 220 peristiwa antara tahun 1850-1920 menggunakan alat penganotasian SAHA digabungkan dengan pelayan perpustakaan Ontologi ONKI untuk menggunakan ontologi domain kongsi. Kesimpulannya, ontologi sejarah menentukan URI untuk peristiwa boleh digunakan untuk menganotasi objek budaya lain dan menghubungkannya dengan satu sama lain. Walau bagaimanapun, ontologi peristiwa tidak dapat diakses oleh orang ramai dan dianggap sebagai ontologi yang tidak boleh diperkembangkan.

c. Ontologi Sejarah FDR

Matlamat utama projek FDR / Pearl Harbour adalah membangunkan aplikasi yang dapat membantu meningkatkan pencarian dan mendapatkan maklumat daripada satu set dokumen yang diambil dari Perpustakaan Presiden Franklin D. Roosevelt (FDRL). Projek ini menggunakan satu set dokumen yang merujuk kepada situasi dan peristiwa

sepanjang tempoh sepuluh tahun iaitu sebelum pengeboman Pearl Harbor. Projek Pelabuhan FDR / Pearl Harbor membina ontologi sejarah berdasarkan model yang dibentangkan menggunakan entiti dan peristiwa dalam pengumpulan dokumennya (Ide&Woolner 2007). Ontologi temporal FDR hanya memasukkan entiti endowmen yang jelas dalam pengumpulan dokumen, yang terdiri daripada kategori umum berikut: entiti geopolitik, organisasi geopolitik, organisasi ketenteraan, kenderaan tentera, objek geografi, artifak geografi, dokumen, perjanjian, orang dan organisasi politik. Penganotasian peristiwa dan entiti dokumen menggunakan *General Architecture for Text Enggineering (GATE)* untuk melengkapkan penganotasian semantik manual. Selanjutnya, penganotasian automatik dijalankan menggunakan pembelajaran mesin berdasarkan anotasi yang disahkan tangan. Bagaimanapun, ontologi temporal FDR tidak dapat diakses oleh orang ramai dan dianggap sebagai ontologi yang tidak boleh diperkembangkan.

d. Ontologi RDF/OWL - *Henry III Fine Rolls*

Henry III adalah projek kerjasama antara King's College London dan National Archives (UK). Tujuan utama projek ini adalah untuk mewakili kerumitan dokumen sejarah yang dikenali sebagai Fine Rolls (Vieira&Ciula 2007). Ontologi FRH3 terdiri daripada beberapa kelas seperti kuasa (Orang, Tempat, dan Subjek) dan *Factoid* (Peranan, Hubungan dan Peranan_Hubungan). RDF/OWL telah dipilih untuk melakukan senarai kuasa berdasarkan beberapa sebab: 1) Ia adalah piawai W3C untuk Web Semantik; 2) Bilangan alat sedia ada adalah lebih tinggi untuk RDF / OWL; 3) Ia boleh dinyatakan sebagai XML, mempermudahkan proses penyampaian data dan ini memudahkan untuk mengindeks orang, tempat dan subjek menggunakan XSLT; 4) Ia boleh mewujudkan ekspresi hubungan antara tika yang dijelaskan dalam bahan sumber Fine Rolls (Vieira&Ciula 2007). Walau bagaimanapun, ontologi ini tidak dapat diakses oleh orang ramai dan dianggap sebagai ontologi yang tidak boleh diperkembangkan.

e. Ontologi Akses kepada Maklumat Muzium

Ontologi Akses kepada maklumat muzium boleh diwakili sebagai "ontologi teras" yang menggabungkan entiti asas dan hubungan merentasi pelbagai kosa kata metadata

(Signore 2005). Ontologi teras berguna dalam membantu mengintegrasikan maklumat dari pelbagai kosa kata dan proses seragam merentas pelbagai sumber maklumat. Ontologi teras adalah model formal teras asas untuk alat yang mengintegrasikan data sumber dan melaksanakan pelbagai fungsi (Signore 2005). Terdapat beberapa kelas dalam ontologi ini seperti *E2 Temporal Entity, E52 Time-span, E3 Condition State, E4 Period* dan *E5 Event*. Proses ontologi juga membantu memperkayakan pengetahuan. Oleh itu, tahap kerumitan yang lebih tinggi boleh diterima dan reka bentuk harus lebih termotivasi oleh ketepatan logik dan kesempurnaan daripada pemahaman manusia. Walau bagaimanapun, ontologi teras ini tidak dapat diakses oleh orang ramai dan dianggap sebagai ontologi yang tidak boleh diperkembangkan.

f. Ontologi *Simple News and Press(*SNaP*)*

Ontologi SNAP adalah ontologi berita yang terdiri daripada pelbagai ontologi, yang menggambarkan *assets (text, image, video), event* serta entiti *(people, places, organizations, abstract concepts dan sebagainya), stuff, tag, classification* dan *identifier*. Terdapat dua kategori entiti dalam ontologi SNaP: entiti mudah iaitu *stuff* dan entiti kompleks iaitu *event*. Istilah *stuff* boleh diwakili sebagai konsep abstrak iaitu *intangible* serta perkara yang nyata iaitu *tangible*. Ontologi *Event* dalam SNaP diwarisi sepenuhnya dari *public domain event ontology*. *Event* merupakan entiti majmuk dalam domain penyelidikan ini (i.e. ia merupakan entiti utama yang mempunyai banyak hubungan dengan entiti lain seperti *people, places, organizations, abstract concepts dan sebagainya.* Jumlah bilangan konsep yang terlibat dalam ontologi *event* dan *stuff* ialah 22 konsep. Walaupun ia bertujuan untuk dokumen berita, ia didapati bersesuaian dalam kajian ini kerana ia mengandungi perwakilan terperinci mengenai *people, organizations, locations, tangible* dan *intangible.* Ontologi SNaP boleh diakses oleh orang ramai dan boleh dianggap sebagai ontologi yang diperkembangkan.

Sebagai kesimpulan, berdasarkan kajian di atas, beberapa ciri penting untuk memilih ontologi yang sesuai untuk diperluaskan telah dikenal pasti. Ciri yang paling penting adalah ketersediaan di mana ontologi sedia ada boleh diakses untuk digunakan semula dan kemudiannya dibangunkan berdasarkan domain khusus. Sebagai contoh,

hanya ontologi STOLE dan SNaP boleh diakses oleh orang ramai. Di samping itu, kita juga perlu mengetahui saiz dan kandungan ontologi untuk memudahkan perluasan ontologi. Sebagai contoh, ontologi SNaP dan STOLE mempunyai bilangan konsep yang paling banyak berbanding dengan ontologi lain. Namun begitu, ontologi STOLE menggunakan bahasa Itali yang mana bahasa ini tidak difahami oleh jurutera ontologi. Justeru itu, sukar bagi jurutera ontologi untuk membina ontologi baru. Dengan ini, hanya ontologi SNaP mempunyai potensi untuk digunakan semula bagi kajian ini. Akhir sekali, kami juga mengkaji jika terdapat alat yang sesuai untuk digunakan dalam melaksanakan proses anotasi. Sebagai contoh, kebanyakan kajian tidak menjelaskan alat anotasi yang sesuai kecuali ontologi sejarah FDR dan ontologi peristiwa. Oleh itu, berdasarkan kajian semua ciri ini, ontologi SNaP mempunyai potensi besar untuk digunakan semula dalam pembinaan ontologi baru. Jadual 2.3 menunjukkan ringkasan ciri ontologi sejarah dan peristiwa sedia.

Jadual 2.3 Ringkasan ciri ontologi sejarah dan peristiwa sedia ada

Ontology	**Jumlah konsep**	**Alatan Anotasi**	**Ketersediaan**	**Bahasa**
Ontologi STOLE	14	Tiada	Ya	Itali
Ontologi Peristiwa	Tiada	SAHA	Tidak	Inggeris
Ontologi Sejarah FDR	10	GATE	Tidak	Inggeris
Ontologi RDF/OWL	8	Tiada	Tidak	Inggeris
Ontologi Akses	5	Tiada	Tidak	Inggeris
Ontologi SNaP	22	Tiada	Ya	Inggeris

2.4 PENDEKATAN ONTOLOGI DALAM DOMAIN SEJARAH

Pendekatan ontologi telah menerima pengiktirafan daripada akademik dan industri dalam pelbagai bidang. Salah satu domain yang mendapat perhatian pada masa kini ialah sejarah. Faktor ini adalah disebabkan pertambahan dokumen sejarah terdigit dan artifak yang boleh diakses oleh orang ramai. Dalam konteks arkib sejarah, Elena et al. (2010) menjelaskan bahawa ahli sejarah menggunakan pengetahuan, pengalaman dan gerak hati mereka untuk membuat keputusan mengenai maklumat yang perlu dicari dan belajar dari sumber yang relevan. Hasil penemuan dari penyelidikan Elena et al. (2010) ini jelas menyatakan bahawa ahli sejarah memerlukan repositori sumber sejarah dan alat pembinaan bagi membolehkan mereka mengakses maklumat secara lebih berkesan dan komprehensif. Antara maklumat penting bagi mereka ialah

peristiwa. Oleh itu, salah satu cara untuk membantu mereka mencapai maklumat dari repositori yang besar ialah sistem capaian maklumat. Capaian maklumat ialah satu bidang yang prihatin terhadap struktur, analisis, pengurusan, penyimpanan, gelintaran dan capaian mengenai maklumat. Menurut Katifori et al. (2016) bidang capaian maklumat telah membawa kepada pembangunan Web Semantik. Walau bagaimanapun, sistem konvensional capaian maklumat tidak dapat menyokong keperluan pembangunan web semantik disebabkan hanya menggunakan perwakilan dokumen *bag-of-words* yang ringkas. Model capaian maklumat *bag-of-words* mewakili dokumen dengan hanya satu struktur berdasarkan set perkataan dan tidak membenarkan permodelan hubungan antara subset kata(Liu et al. 2007). Oleh yang demikian, salah satu cara yang dapat membantu menyokong keperluan tersebut ialah dengan menggunakan pendekatan ontologi. Definifi ontologi itu sendiri memfokuskan pada aspek hubungan antara konsep dalam pembinaan ontologi. Ia turut berfungsi meningkatkan perwakilan dokumen secara semantik bagi capaian maklummat yang lebih tepat. Beberapa kajian lepas berkenaan pembangunan ontologi dalam domain sejarah yang melibatkan elemen peristiwa dikaji bagi mengetahui kepentingan elemen peristiwa kepada pembangunan ontologi serta kewujudan aspek hubungan antara konsep dalam ontologi.

Kajian Corda (2007) menggunakan teori Davidson semasa membangunkan model ontologi untuk sejarah Sains. Model ini memperkenalkan teknik pengelasan masa dan kaedah untuk mewakili masa dalam hubungan konsep. Masa boleh diwakili dalam beberapa dimensi seperti konsep masa, hubungan tempoh masa dan hubungan yang terjadi dalam sesuatu peristiwa. Dimensi pertama adalah kelas atas yang mengisytiharkan spesifikasi tempoh masa yang utama. Seterusnya, dimensi kedua dan ketiga menggunakan teori Davidson secara berasingan di kedua-dua peringkat model data, yang terdiri daripada hubungan tempoh masa dan hubungan dengan serta-merta. Selain daripada itu, Corda (2007) telah memperkenalkan pendekatan awal untuk model masa yang meliputi ketahanan, ketepatan masa dan kejadian berulang serta kejadian bukan berulang. Di samping itu, sumbangan utama Corda (2007) adalah menggabungkan teori Davidson mengenai peristiwa dan Aljabar selang Allen ke dalam satu rangka kerja, terutamanya untuk penambahan dan perbandingan masa

sesuatu peristiwa. Gabungan ini telah memastikan hasil yang konsisten dan ia kelihatan kukuh sebagai asas untuk peningkatan selanjutnya.

Selain itu, IdeandWoolner (2007) telah membangunkan ontologi sejarah dan ditakrifkan sebagai H = <OS, OE, T>, iaitu OS disebut sebagai siri ontologi temporal, OE disebut sebagai ontologi masa dan peristiwa dan T disebut jangka masa. Ontologi masa dan peristiwa terdiri daripada OE = <R,T>, yang mana R adalah satu set tika yang mewakili peristiwa berlaku dalam sesuatu masa seperti serangan ke atas Pearl Harbor, komunikasi di antara Roosevelt dan setiausaha negara, pengenaan sekatan minyak ke atas Jepun dan sebagainya, dan selang temporal bagi setiap peristiwa. T mewakili keseluruhan jangka masa dalam siri ontologi temporal. Tika peristiwa adalah saling berkaitan dengan selang temporal.

Di samping itu, IdeandWoolner (2004) menyatakan bahawa dalam penyelidikan sejarah, mengenalpasti pelbagai 'peristiwa' dalam data adalah penting. Peristiwa sejarah dikategorikan kepada dua kelas seperti $KeyEvents$ dan *InformationalEvents*. Contoh *KeyEvents* adalah seperti pencerobohan tentera atau perubahan dalam pentadbiran, yang menyebabkan pengubahsuaian dalam ontologi dan dikaitkan dengan jarak masa pada titik perubahan. Sementara itu, *InformationalEvents* boleh dilihat dalam contoh yang lain seperti melawat ke Rumah Putih oleh Duta Jepun yang mana tiada perubahan terhadap ontologi. Selain daripada itu, terdapat beberapa peristiwa sejarah yang tidak pernah disebut dalam mana-mana dokumen sejarah yang mana ia sebenarnya merupakan sebahagian daripada pengetahuan am sejarah yang tidah kurang pentingnya kepada masyarakat. Contohnya seperti peristiwa besar yang terjadi sebelum dan semasa Perang Dunia II. Kedua-dua kelas peristiwa sejarah *KeyEvents* dan *InformationalEvents* boleh dipetakan melalui penyatuan, pengasingan, perubahan nama, dan lain-lain. Kelas pemetaan dikenali sebagai *OntologyChangeEvent*. Kelas *OntologyChangeEvent* telah dibangunkan berdasarkan konsep GrenonandSmith (2004), yang membolehkan penalaran untuk mengubah peristiwa itu sendiri. Sebagai contoh, kelas *OntologyChangeEvent* boleh digunakan untuk mencorak punca sesuatu hubungan mengikut kehendak tertentu ahli sejarah. Walaupun pembangunan ontologi sejarah FDR memberikan beberapa sumbangan yang paling besar bagi masyarakat, tetapi ia menimbulkan beberapa

cabaran yang menarik mengenai pengkonteksan temporal. Projek FDR atau Pearl Harbor membangunkan satu ontologi sejarah dan ia merupakan lanjutan daripada *Suggested Upper Merged Ontology* (SUMO), *Mid-Level Ontology* (MILO) dan beberapa ontology sedia ada dari *Agent Semantic Communication Service* (ASCS). Projek ini bertujuan menyokong penambahbaikan gelintaran dan capaian dokumen dari *Franklin D. Roosevelt Presidential Library* (FDRL).

Vossen et al. (2009) memfokuskan kepada kajian mengenai pembinaan ontologi sejarah dan kamus, yang mana sumber tersebut akan digunakan dalam sistem capaian maklumat terkini. Sistem terkini yang dicadangkan mampu menangani isu dinamik berasaskan masa dan perspektif yang berbeza dalam arkib sejarah. Vossen et al. (2009) menerangkan bahawa berita sedia ada akan menjadi sejarah pada masa akan datang. Sebagai contoh, pembinaan sesuatu ayat mempegaruhi definisi sejarah sepertimana ayat “seseorang menembak presiden” merujuk kepada berita bukannya sejarah. Manakala ayat “pembunuhan presiden” pula merujuk kepada sejarah bukannya berita. Dalam kes berita, peralihan beransur-ansur daripada laporan berita menjadi rujukan sejarah merupakan isu utama dalam penyelidikan sejarah. Malah isu ini bukan sahaja memberi kesan tetapi juga bagaimana peristiwa tersebut dikonsepkan. Ia akan menjadi lebih jelas sekiranya terdapat banyak cara untuk merujuk kepada peristiwa yang sama dari perspektif sejarah. Oleh itu, pangkalan data leksikal dibina bagi menyediakan pendekatan pemetaan dari bahasa kepada makna bagi pangkalan data sejarah dalam bahasa Belanda dan Inggeris. Seterusnya, kosa kata berbilang bahasa dan ungkapan dalam pangkalan data sejarah dihubungkan kepada ontologi sejarah yang mana sama seperti kajian Ide dan Woolner. Model ontologi ini digunakan oleh perisian pelombongan teks bagi mengekstrak metadata yang dikehendaki dari pangkalan data dan kemudiannya menghasilkan satu indeks semantik. Di akhir kajian ini, model ontologi ini digabungkan kepada indeks semantik.

IdeandWoolner (2007) juga membincangkan isu dinamik yang berasaskan masa dan menyatakan bahawa sejarah boleh ditakrifkan sebagai rekod realiti yang berbeza dari segi masa serta secara khusus memberi tumpuan kepada perubahan dalam realiti. Pada tahun 2010, CybulskaandVossen (2010) membangunkan projek

Semantic of History bagi memfokuskan kepada perubahan realiti yang sentiasa berubah-ubah dari masa ke semasa serta pelbagai sikap penulis teks sejarah terhadap realiti yang sentiasa berubah-ubah. Kedua-dua kerja penyelidikan ini menyumbang kepada peranan IR dalam sejarah. CybulskaandVossen (2010) membincangkan perihal peristiwa sejarah yang direalisasikan dalam pelbagai jenis teks dan apakah implikasi mengenai permodelan maklumat peristiwa. Dalam bidang sejarah, terdapat ramai penulis yang mempunyai perspektif berbeza dari genre yang berbeza dalam pelbagai penggunaan bahasa. Oleh yang demikian, pembangunan model peristiwa penting bagi menunjukkan perbezaan perwakilan peristiwa dalam pelbagai jenis teks dan mengenalpasti hubungan di antara peristiwa sejarah. Kepelbagaian penggunaan bahasa ini menyumbang kepada permasalahan dalam sistem gelintaran bagi mendapatkan maklumat secara semantik berdasarkan kueri pengguna. Oleh itu, kekerapan dalam penggunaan bahasa merupakan penyelesaian bagi mengautomasikan capaian maklumat dari artikel berita dan teks sejarah. CybulskaandVossen (2010) membina kamus dan menjalankan analisis terhadap kamus tersebut bagi memperlihatkan perbezaan terhadap kekurangan teks dari perspektif sejarah yang mana spesifik kepada individu, masa dan tempat.

Hyvönen (2009) membangunkan portal semantik untuk warisan budaya. Terdapat pelbagai jenis kandungan budaya di laman web (dokumen, imej, trek audio, item koleksi, objek pembelajaran dan lain-lain), terdiri daripada beberapa topik (seni, sejarah, kraftangan, dan sebagainya) yang ditulis dalam bahasa yang berbeza dan disasarkan untuk pengguna baru dan pakar-pakar terutamanya yang disediakan oleh organisasi yang berbeza (muzium, arkib, perpustakaan) dan individu. Kesukaran untuk mencari dan mengaitkan maklumat dalam pelbagai bentuk penyediaan kandungan dan persekitaran format data adalah cabaran utama dalam kerja-kerja penyelidikan ini. Ia adalah satu cabaran kepada organisasi dalam menghasilkan kandungan.

Dengan portal semantik, kesukaran untuk mencari maklumat yang berkaitan boleh diselesaikan dengan mengumpul pelbagai kandungan penerbit ke dalam laman tunggal. Terdapat pelbagai jenis portal termasuk portal perkhidmatan (contohnya, Yahoo! dan halaman permulaan yang lain), portal komuniti dan portal maklumat.

Sementara itu, kebanyakan kandungan laman web semantik akan diterbit menggunakan portal maklumat semantik. Portal tersebut berdasarkan piawai web semantik dan kandungan difahami mesin, iaitu, metadata, ontologi dan peraturan. Aplikasi ini mentakrifkan web semantik "model kek lapis" oleh Tim Berners-Lee sebagai model kandungan untuk portal budaya semantik. Ia membuat perbezaan di antara tahap data sintaksis berdasarkan tahap XML dan semantik seperti tahap metadata, tahap ontologi, tahap logik dan tahap kepercayaan.

Hyvönen (2009) mencadangkan satu penyelidikan, iaitu aras ontologi diperkenalkan dengan mewakili ontologi menggunakan Skema RDF dan OWL. Penyelidikan ini menerangkan perbendaharaan kata dan konsep yang berkaitan dalam dunia sebenar. Sebagai contoh, Portal MuseumFinland menggunakan ontologi berkongsi untuk mengisi nilai-nilai unsur metadata. Selepas suatu masa yang tertentu, banyak tesaurus budaya telah berubah menjadi format SKOS. SKOS adalah model data yang berfungsi untuk berkongsi dan menghubungkan sistem organisasi pengetahuan[1]. Dari sudut semantik, transformasi tidak begitu mencukupi kerana versi SKOS sentiasa mengelirukan komputer, terutama dalam pencarian dari segi pengindeksan dan memahami hubungan yang mana memerlukan pengetahuan manusia secara tersirat. Oleh itu, struktur semantik tesaurus ditapis dan disusun semula ke dalam ontologi light-weight sebagai penyelesaian kepada isu ini. Terdapat beberapa domain ontologi digunakan bagi menerangkan metadata budaya. Walau bagaimanapun, situasi ini membawa kepada isu yang berkaitan dengan pembangunan ontologi saling kendali(ontologies mutually interoperable) berlaku. Isu ini timbul semasa berurusan dengan kandungan yang sepadan dengan skema metadata yang berbeza dalam setiap ontologi. Oleh itu, pemetaan dan penjajaran ontologi dicadangkan untuk berhadapan dengan isu ini. Sebagai contoh, Hyvönen (2009) mencadangkan CULTURESAMPO, yang merupakan satu kajian mengenai masalah saling kendali semantik berkenaan skema metadata pada peringkat perbendaharaan kata domain secara lebih mendalam. Isu ini perlu diselesaikan kerana kini kandungan budaya boleh didapati dalam pelbagai bentuk seperti objek budaya, proses kebudayaan, bangunan kebudayaan dan laman web, dan peristiwa-peristiwa sejarah.

[1] https://www.w3.org/TR/skos-reference/

Oleh kerana isu ini, integrasi kandungan dilakukan dengan mengubah kandungan ke dalam skema perwakilan pengetahuan ontologi ringan(*light-weight knowledge representation scheme*) berdasarkan domain ontologi peristiwa dan peranan termatik. Ontologi berdasarkan klasifikasi temporal peristiwa pada garis masa boleh digunakan untuk mengelolakan peristiwa-peristiwa yang berbeza. Selain itu, peristiwa juga boleh diklasifikasikan sebagai jenis peristiwa atau cabang sejarah(Hyvönen 2009).

Pada tahun 2012, Corda et al. (2012) menyatakan carian maklumat dalam domain sejarah adalah kompleks, mempunyai pelbagai elemen dan memerlukan penyatuan serta penerokaan terhadap hubungan di antara peristiwa. Ini adalah kerana kurangnya perdekatan formal untuk membina rangka kerja bagi menghubungkan peristiwa sejarah. Oleh sebab itu, penyelidikan ini memfokuskan kepada aspek penggabungan perwakilan ontologi peristiwa dan satu model sistematik bagi membina hubungan di antara peristiwa(Trajektori Semantik). Satu pendekatan formal bagi mekanisma Trajektori Semantik dihasilkan untuk membantu pengguna mendapatkan idea utama dan hubungan penting yang berkaitan kepada carian maklumat pengguna. Pendekatan formal ini diambil dari Ontologi Peristiwa dan elemen semantiknya diperkaya oleh peraturan serta hubungan definisi templat.

Seterusnya pada tahun 2014, Afolabi et al. (2014) membangunkan ontologi domain untuk sejarah Nigeria. Pembangunan ontologi ini adalah disebabkan faktor terlalu banyak dokumen sejarah yang disimpan dalam bentuk cetakan seperti buku sejarah, laporan media dan artifak. Matlamat pembangunan ontologi domain ini ialah memberi sokongan kepada keperluan automasi runcit yang memerlukan pengetahuan sejarah dengan membina satu pangkalan pengetahuan peristiwa sejarah. Selain itu, pembangunan ontologi ini turut menghasilkan satu ontologi piawai yang membolehkan perkongsian konsep sejarah Nigeria dalam sesuatu komuniti serta sistem. Ontologi domain ini dibina menggunakan pendekatan separa automatik yang melibatkan pra-pemprosesan kata dari sumber teks dan akhirnya pengkonsepan serta permodelan. Pembangunan ontologi domain menggunakan alatan *Protege Ontology Editor*.

Kesimpulannya, berdasarkan ulasan di atas terdapat banyak kajian telah dilakukan dalam bidang sejarah yang mengaplikasikan pendekatan ontologi pada elemen peristiwa. Semua kajian mengaplikasikan model ontologi bagi membolehkan pengkongsian konsep dengan sifat hubungan yang berkaitan bagi mentafsirkan idea utama dan hubungan penting yang berkaitan kepada carian maklumat pengguna. Oleh itu, pendekatan ontologi merupakan satu kaedah sistematik yang dapat membantu bidang sejarah menyelesaikan masalah yang timbul dalam perwakilan dan perkongsian maklumat secara lebih spesifik terutamanya dalam elemen peristiwa. Seperti mana yang telah dinyatakan oleh (Joseph&Janda 2008) bahawa sejarah sering memberi tumpuan kepada peristiwa dan perkembangan yang berlaku pada masa tertentu. Oleh yang demikian, elemen peristiwa dalam domain sejarah dipilih sebagai fokus utama kepada penyelidikan ini.

2.5 CAPAIAN MAKLUMAT

Capaian maklumat didefinisikan sebagai menyimpan, menyusun, mewakili, mencari dan mencapai maklumat yang sepadan dengan permintaan pengguna(Korfhage 1997). Pendekatan konvensional capaian maklumat bermula dengan maklumat teks dan kebanyakannya dicadangkan untuk menguruskan sejumlah besar maklumat berasal dari bidang sains perpustakaan(Alidin 2007; Ren&Bracewell 2009; Sanderson&Croft 2012).

Bush (1945) menerbitkan sebuah artikel tentang *As We May Think* yang mencadangkan akses automatik untuk sejumlah besar pengetahuan yang disimpan. Dengan terhasilnya idea ini, bidang capaian maklumat dilahirkan pada tahun 1950. Dalam pertengahan tahun 1950-an, beberapa penyelidikan capaian maklumat muncul dengan kewujudan komputer untuk pencarian teks. Penyelidikan capaian maklumat berkembang pesat kerana peningkatan pelbagai teknologi komputer seperti penambahan kelajuan pemprosesan dan kapasiti simpanan. Pada mulanya penyelidikan capaian maklumat hanya berkisar tentang pencarian maklumat yang berkaitan dengan permintaan pengguna. Satu sistem capaian maklumat biasanya dibina untuk data berstruktur dan separa berstruktur (i.e.: laman web, dokumen, imej, video dan lain-lain)(Sanderson&Croft 2012). Sistem ini diperlukan terutamanya

apabila saiz maklumat yang meningkat sehingga teknik pengkatalogan tradisional tidak lagi dapat menampung penyimpanan dan pencarian maklumat(Walter 2005).

Bermula pada tahun 1960-an, Gerard Salton telah membentuk dan mengetuai sekumpulan penyelidik capaian maklumat yang besar. Kumpulan ini telah mencipta satu idea dan konsep dalam bidang penyelidikan seperti formalisme algoritma untuk melaksanakan proses pemangkatan dokumen relatif bagi sesuatu kueri (Sanderson&Croft 2012). Keperluan sistem capaian maklumat dalam proses pemangkatan capaian terus berkembang sehingga tahun 1970-an. Sehingga kini, bidang kajian ini masih menjadi tumpuan utama dalam bidang penyelidikan. Sebagai contoh, Jones et al. (2001) menyusun kedudukan setiap maklumat pengguna berdasarkan kaedah kedekatan geografi iaitu berdasarkan nama tempat dan khususnya mengenalpasti topik yang diminati berdasarkan pendekatan semantik.

Dalam tahun 1980-an hingga pertengahan 1990-an, skop capaian maklumat bukan sahaja memberi tumpuan kepada kueri pengguna dalam bidang sains perpustakaan tetapi juga berkembang kepada model formal mengenai capaian maklumat. Capaian maklumat juga berfungsi bagi membantu masyarakat selaras dengan perkembangan teknologi Internet dan *World Wide Web* yang telah diperkenalkan oleh Berners-LeeandFischetti (2001). Capaian maklumat telah menjadi satu keperluan kepada masyarakat apabila ia mula digunakan secara meluas setiap hari dalam pelayar web seperti *Yahoo*, *Google* atau beberapa sistem yang khusus dicipta untuk perpustakaan. Pada pertengahan tahun 1990-an hingga kini kemunculan enjin carian web adalah sejajar dengan pertumbuhan *World Wide Web*.

2.5.1 Senibina dan komponen capaian maklumat

Sistem capaian maklumat mempunyai empat komponen asas iaitu pengumpulan dokumen, pengindeksan, pencarian dan pengurusan dokumen dan kueri.

a. Pengumpulan dokumen

Fasa ini menjelaskan proses IR dihasilkan sebagai output pembinaan dan penyelenggaraan dari masa ke semasa untuk koleksi dokumen maya dan fizikal.

Objektif fasa ini ialah membina koleksi dokumen yang dapat menarik alatan perisian untuk pengindeksan dan pengurusan serta set dokumen ini juga adalah yang terbaik untuk diberikan kepada pengguna.

b. Pengindeksan

Fasa ini menerangkan proses IR yang mana mengambil satu dokumen dari koleksi sebagai input dan mewakilkan kandungan tersebut untuk menjadikan ia boleh digunakan oleh satu proses komputer. Proses ini boleh dijalankan dengan menggunakan pendekatan yang berbeza: 1) satu pendekatan pengindeksan automatik yang lengkap mengenai dokumen teks penuh - automatik 2) Sesuatu dokumen bagi satu kelas hirarki pengetahuan yang dibina dan diselenggara oleh manusia - semi-automatik.

c. Pencarian

Fasa ini menguruskan kueri pengguna menggunakan algoritma IR dan bertujuan memilih dokumen yang paling relevan dengan kueri pengguna. Fungsi carian adalah berkait rapat dengan algoritma indeks yang mana telah digunakan dalam perwakilan kandungan dokumen.

d. Pengurusan dokumen dan kueri

Sebahagian dari fasa ini adalah bertindih dengan fasa pengindeksan dan gelintaran kerana pengurusan dokumen adalah berkait dengan pengindeksan dan pengurusan kueri untuk gelintaran. Kebiasaannya fasa ini diperlukan sebagai fasa berasingan iaitu merangkumi isu storan dan pengurusan indeks yang terhasil dari proses indeks serta mempunyai pangkalan data berlainan yang berguna untuk mengurus dan menawarkan khidmat yang berbeza kepada pengguna.

2.5.2 Model capaian maklumat

Sistem capaian maklumat boleh dikategorikan kepada tiga kumpulan utama dari segi model yang digunakan untuk perwakilan dokumen dan kueri yang mempunyai pengaruh yang besar terhadap prestasi capaian.

a. Model Boolean

Model ini merupakan satu model mudah berdasarkan teori set dan aljabar Boolean (Heaps 1978). Dokumen diwakili sebagai set istilah dan kueri bagi menyatakan ungkapan Boolean. Dokumen dicapai jika ia mengandungi semua istilah yang dinyatakan oleh ungkapan Boolean. Walaupun ia intuitif dan mudah dilaksanakan, namun ia memerlukan padanan tepat dalam capaian, yang menyebabkan masalah "terlalu banyak atau tiada".

b. Model Ruang Vektor

Model ini mewakili kedua-dua dokumen dan kueri sebagai vektor istilah berpemberat. Dokumen dicapai mengikut tahap kesamaan dengan vektor kueri. Berbeza dengan model Boolean, model ruang vektor juga boleh mencapai dokumen yang dipadankan secara sebahagian. Model ini dicipta oleh G. Salton (Salton et al. 1975).

c. Model Kebarangkalian

Dalam model ini, satu set dokumen yang relevan untuk kueri diandaikan. Capaian dilakukan mengikut kebarangkalian kepunyaan set ini. Kebarangkalian biasanya diperolehi menggunakan Teorem Bayes (Van Rijsbergen 1979).

Setelah menjelaskan model IR yang berlainan, kajian ini secara ringkas menyatakan mengenai model IR yang digunakan dalam capaian maklumat berasaskan ontologi. Model ontologi ini menggunakan model VSM untuk capaian. Penggunaan model VSM telah menjadi perhatian penyelidik pada masa kini kerana ia merupakan satu paradigm perwakilan mudah bagi mewakili dokumen dan kueri(Styltsvig 2006). Pada dasarnya, VSM digunakan untuk mendapatkan dokumen yang berkaitan dengan

mengira frekuensi istilah(tf) dan frekuensi dokumen songsang(idf) seperti yang ditunjukkan dalam dalam Bab 4.

2.5.3 Proses Asas Capaian Dokumen

Menurut Roitblat (2001), kaedah yang sering digunakan bagi capaian dokumen ialah carian kata kunci. Dengan menggunakan kaedah ini, pengguna memasukkan kueri yang terdiri daripada satu atau beberapa perkataan dan seterusnya sistem akan memaparkan semua dokumen yang telah diindeks dan mengandungi perkataan tersebut. Akhirnya, proses pemangkatan dijalankan untuk mendapatkan dokumen yang bertepatan dengan kueri pengguna. Proses pemangkatan ini menggunakan kaedah pengiraan pemberat bagi setiap dokumen yang terlibat serta ukuran kesamaan antara kueri pengguna dan dokumen.

2.5.4 Masalah utama capaian dokumen

Ada kalanya dokumen yang dicapai melalui carian piawai tidak relevan dengan permintaan pengguna. Ini adalah kerana capaian dokumen menggunakan kaedah ini hanya tertumpu kepada carian sesuatu istilah atau pun kata kunci yang mana tidak menepati maksud dan kehendak pengguna. Masalah ini menjadi lebih rumit apabila jaringan sejagat sentiasa padat dengan jumlah maklumat yang terus meningkat. Oleh yang demikian, pencarian maklumat yang relevan dengan kehendak pengguna menjadi satu cabaran baru pada hari ini. Sebagai contoh, pengguna hendak mencari sub peristiwa dan peristiwa berkaitan kepada sesuatu peristiwa A tetapi sistem tidak dapat mencapai dokumen yang berada dalam kategori yang berbeza (iaitu sub peristiwa dan peristiwa berkaitan). Masalah ini timbul kerana terdapat batasan utama mengenai pemahaman terhadap sesuatu maklumat yang mana tidak dapat menjelaskan maksud maklumat tersebut secara terperinci kepada pengguna. Dengan adanya teknologi semantik, penggunaan tika dalam konsep relevan untuk dokumen semantik hendaklah dinyatakan dengan jelas dalam sesuatu proses bagi mewakili data. Oleh itu, pembangunan model baru secara menyeluruh membantu menyelesaikan proses perwakilan data dengan betul dan jelas.

2.6 CAPAIAN MAKLUMAT DOKUMEN SEJARAH

Bidang capaian maklumat terus berkembang sejajar dengan perubahan persekitaran pengkomputeran. Sebagai contoh, pertumbuhan pesat penyelidikan dalam capaian maklumat dokumen sejarah membolehkan capaian sesuatu dokumen sejarah dibuat mengikut kueri pengguna secara lebih tepat. Dokumen sejarah boleh ditakrifkan sebagai penyimpanan maklumat berkaitan dengan ruang masa yang mana dokumen yang diterbitkan pada masa lampau masih boleh digunakan pada masa akan datang (Cabo&Llavori 1998).

Satu kajian oleh Jones et al. (2001) lebih memberi tumpuan kepada capaian dokumen dalam konteks geografi. Dalam penyelidikan ini, IR menyokong konteks geografi dalam proses pencarian dokumen, imej atau rekod yang mana bergantung kepada ruang geografi sahaja iaitu dengan menggunakan kueri nama sesuatu tempat. Isu timbul apabila nama kueri dan nama yang berkaitan dengan sumber maklumat pengguna mengalami padanan tidak tepat. Ini disebabkan oleh kata klasifikasi bukan spatial yang mana tidak boleh diakses terus melalui kaedah pengindeksan berdasarkan koordinat. Oleh itu, capaian maklumat geografi dihasilkan untuk menyusun kedudukan maklumat pengguna yang relevan dengan kawasan geografi yang berdekatan serta ketepatan semantik berkaitan dengan topik yang menarik. Sebagai contoh, data yang dilaksanakan termasuk dokumen teks moden dan sejarah (semua rekod peristiwa kebudayaan dan alam sekitar serta perihal bahan disimpan di muzium i.e.: institusi penyelidikan dan repositori arkib lain). Bagi menyelesaikan isu ini, ontologi tempat dibentangkan yang mana menggabungkan data selaras terhad dengan hubungan spatial kualitatif di antara tempat. Ontologi menggunakan pautan sistem pemodelan semantik kepada hierarki konseptual bukan spatial. Dengan adanya langkah ini, maka kaedah untuk memadankan nama tempat tertentu dengan nama tempat yang merujuk kepada lokasi berhampiran dihasilkan. Kajian ini memberi sumbangan yang cukup besar dalam mempelbagaikan jenis maklumat yang diselenggarakan dan membangunkan langkah ketepatan semantik untuk mengautomasikan susunan bagi keputusan sesuatu kueri.

Seterusnya pada awal abad ke-20 dan ke-21 telah berlaku perubahan terhadap corak masyarakat dalam mengakses maklumat. Pengguna menjangkakan banyak maklumat sejarah yang boleh dikongsikan dan digunakan semula melalui perpustakaan digital, yang mana boleh memberi padanan carian dokumen yang lebih tepat dengan menggunakan sistem kecekapan menjawab soalan di samping memberi sokongan kepada senario yang terpilih (Corda 2007; Mirzaee et al. 2004). Bagi memenuhi permintaan pengguna, Mirzaee et al. (2004) dan Corda (2007) mencadangkan aplikasi semantik dalam dokumen sejarah yang mana ia membantu memperkayakan capaian, penggunaan semula dan manipulasi mengenai pengetahuan terbenam yang perlu digali. Aplikasi semantik ini memberi hasil yang lebih berkesan dengan pendefinisian hubungan berasaskan masa.

Sejajar dengan pertumbuhan perpustakaan digital, Koolen et al. (2006) turut menjalankan penyelidikan mengenai warisan budaya yang dipelihara dalam perpustakaan, arkib dan muzium yang mana ditulis oleh banyak negara. Kajian ini menyatakan bahawa inisiatif pendigitalan membolehkan pengguna amatur mendapatkan semua dokumen melalui perpustakaan digital dan enjin carian vertikal. Walaubagaimanapun, keputusannya tetap tidak memuaskan, iaitu kueri yang menggunakan perkataan moden mungkin tidak sepadan untuk mencapai dokumen yang mengandungi istilah dalam dokumen sejarah. Justeru itu, pendekatan bahasa silang untuk capaian dokumen bersejarah telah dicadangkan dan seterusnya pembinaan sumber terjemahan automatik untuk bahasa bersejarah telah dikaji oleh Koolen et al. (2006). Hasilnya, capaian dokumen sejarah menggunakan teknik capaian maklumat bahasa silang menunjukkan hasil yang positif dalam menyelesaikan isu yang berkaitan dengan capaian. Pertamanya, Koolen et al. (2006) telah menemui perbandingan yang melibatkan (a) persamaan turutan fonetik, (b) kekerapan relatif urutan konsonan dan vokal dan (c) kekerapan relatif jujukan aksara n-gram bagi korpora sejarah dan moden yang mana boleh membina kaedah bagi memodenkan bahasa bersejarah. Keduanya, penggunaan kueri moden tidak memberikan kesan yang efektif dalam capaian dokumen sejarah berbanding penggunaan beberapa alatan bahasa bersejarah yang banyak memberi peningkatan dalam keberkesanan capaian dokumen. Kesimpulannya, pendekatan bahasa silang boleh membantu merapatkan

jurang antara bahasa bersejarah dalam sesuatu dokumen dan bahasa moden daripada kueri pengguna.

Pilz et al. (2006) membangunkan alat berasaskan web mengenai warisan budaya dan memberi tumpuan kepada carian berasaskan peraturan dalam pangkalan data teks dengan ejaan yang tidak piawai. Kajian ini menghasilkan enjin carian kabur berasaskan peraturan yang membolehkan pengguna untuk mencapai data teks bebas daripada realisasi ortografikal itu. Semua peraturan ini telah dihasilkan oleh analisis statistik, penerbitan sejarah, prinsip linguistik dan pengetahuan pakar. Pilz et al. (2006) juga menggunakan satu penerbitan peraturan automatik dan klasifikasi hasil yang lebih halus melalui sukatan persamaan Levenshtein yang umum bagi fungsi selanjutnya. Enjin carian dalam talian dengan kemudahan penggunaan elemen telah disediakan sebagai sebahagian daripada alat berasaskan web untuk ahli bahasa profesional dan amatur yang berminat.

Di samping itu, Hauser et al. (2007) juga mengkaji tentang warisan budaya yang melibatkan dokumen tersembunyi seperti buku dan dokumen sejarah. Baru-baru ini, kajian mengenai dokumen tersembunyi mendapat banyak perhatian. Dalam usaha untuk mengekalkan dokumen-dokumen ini, mereka perlu didigitalkan. Setelah dokumen didigitalkan, teknik moden akses maklumat seperti capaian maklumat, perlombongan teks, hiperpautan dan persembahan fleksibel mengenai dokumen boleh digunakan. Satu masalah serius timbul apabila capaian kepada dokumen sejarah tidak berjaya disebabkan sejumlah besar variasi ejaan perkataan yang sama tidak boleh menggunakan teknik pengindeksan piawai secara langsung untuk capaian maklumat. Oleh itu, kajian ini memberi perhatian kepada alternatif penyelesaian lakaran untuk masalah ini. Dalam penyelesaian pertama, satu kamus khas digunakan bagi menyimpan setiap kemasukan perkataan moden dalam satu senarai pemantauan varian sejarah. Dalam penyelesaian kedua, penggunaan padanan generatif berasaskan peraturan boleh menggambarkan perbezaan di antara varian ejaan sejarah yang baru dan sepadan dengan membangunkan satu set petua. Akhir sekali, padanan dengan berdasarkan perkataan persamaan diaplikasi bagi mengetahui keserasian varian ejaan baru dan lama.Selain itu, Schockaert et al. (2010), Mirzaee et al. (2005) dan Corda (2007) turut mencadangkan supaya dokumen patut diisih mengikut aspek temporal

untuk meningkatkan sistem capaian maklumat. Alonso et al. (2007) juga menyokong aspek temporal sebagai merupakan satu ciri penting dalam capaian maklumat bagi membantu meningkatkan fungsi aplikasi carian.

Maklumat temporal yang berstruktur diperlukan untuk menyokong kekangan pada capaian maklumat yang menyebabkan ia tidak dapat melaksanakan tugasnya dengan sempurna sebagaimana yang dikehendaki pengguna dan sempadan temporal mengenai kebanyakan peristiwa sejarah adalah tidak jelas. Kekangan temporal telah digunakan sebagai rumusan mencari maklumat tentang peristiwa sejarah dan kueri dalam kajian Schockaert et al. (2010). Selain itu, Schockaert et al. (2010) mencadangkan satu rangka kerja berdasarkan pengaburan daripada Aljabar selang Allen untuk menyelesaikan masalah ini. Teknik heuristik mudah telah digunakan untuk mendapatkan maklumat temporal dari dokumen web dan memberi tumpuan kepada perolehan kembali yang tepat. Tambahan pula, pergantungan pada pentaakul temporal kabur boleh meningkatkan kebolehpercayaan pra-pemprosesan maklumat dan mampu menangani konflik berkaitan kekaburan peristiwa. Hasil kajian menunjukkan bahawa pengetahuan yang konsisten dan boleh dipercayai berdasarkan hubungan temporal kabur boleh diperolehi. Sebagai contoh, baru-baru ini terdapat satu penyelidikan yang mana telah menjawab beberapa soalan temporal terhad pada soalan seperti berikut iaitu Berapa banyak lukisan yang Piet Mondriaan lukis dalam tahun Amsterdam itu?; Di manakah Olimpik Musim Sejuk diadakan sebelum di Salt Lake City? dan Apakah penemuan dan ciptaan baru yang berlaku semasa Revolusi Astronomi? Di manakah dan siapakah yang menemui dan menciptanya? Peristiwa manakah yang mendahului atau berjaya dalam penemuan atau ciptaan?(Corda 2007; Schockaert et al. 2010). Dalam konteks rumusan pelbagai dokumen, maklumat temporal telah digunakan untuk mendapatkan urutan kronologi ayat dari dokumen yang berbeza, untuk meringkaskan maklumat yang relevan tentang peristiwa dari rentetan berita, dan untuk menjana secara automatik gambaran keseluruhan garis masa yang mengandungi peristiwa yang paling penting dari corpus berita(Schockaert et al. 2010). Hasil kajian mengesahkan bahawa teknik heuristik boleh mendapatkan hubungan temporal kualitatif sebagai pengganti kepada tempoh masa yang hilang. Di samping itu, algoritma pentaakul temporal kabur turut digunakan untuk menghapuskan maklumat yang salah daripada hubungan temporal yang diekstrak.

Selain itu, isu ejaan turut mendapat perhatian penyelidik yang mana pengguna menghendaki kata kunci moden boleh dipadankan dengan unsur perkataan ataupun ejaan yang terdapat dalam dokumen sejarah (Gotscharek et al. 2011; Koolen et al. 2006; Pilz et al. 2006). Antara isu tersebut ialah bagaimana untuk mengurangkan jurang perbezaan ejaan teks lama dan moden bagi tujuan kueri pengguna. Oleh itu, beberapa pendekatan telah diperkenalkan seperti pendekatan bahasa silang dan mengembangkan pendekatan berasaskan peraturan bagi merapatkan jurang tersebut (Koolen et al. 2006; Pilz et al. 2006). Seterusnya, menurut Gotscharek et al. (2011) terdapat banyak buku yang disimpan dalam perpustakaan tanpa mengetahui kewujudan maklumat tersebut, kecuali hanya boleh dicapai oleh sekumpulan kecil pakar. Oleh itu, ia tidak memadai dengan hanya mencapai semua dokumen melalui kewujudan dan penyimpanan dokumen dalam bentuk imej digital. Ia boleh menjadi lebih menarik jika dokumen itu boleh didapati dalam bentuk teks dan mampu untuk mendapatkan dokumen itu melalui aktiviti carian di web. Sebagai contoh, satu kajian Perpustakaan Negara Australia melaporkan bahawa jumlah pengunjung telah meningkat apabila koleksinya boleh dicapai melalui indeks gelintaran teks(Holley 2009). Jumlah dokumen yang terlalu banyak ini mempunyai banyak varian ejaan terutamanya dalam dokumen sejarah (Gotscharek et al. 2011). Dengan itu, satu kajian telah dijalankan menggunakan kaedah capaian maklumat tetapi malangnya, kaedah ini gagal untuk menghasilkan keputusan yang memuaskan pada dokumen sejarah. Dalam usaha untuk menyelesaikan isu ini, perkataan moden digunakan sebagai kueri yang mana mempunyai kaitan dengan varian sejarah dalam dokumen yang disyorkan. Kedua-dua cara alternatif untuk menyelesaikan masalah ini adalah melalui prosedur padanan khas dan leksikon untuk istilah dalam dokumen sejarah. Pembinaan leksikon dapat membantu memetakan lema moden kepada bentuk lema moden yang lengkap yang mana prosedur padanan ini boleh digunakan secara dalam talian. Sebagai contoh, apabila pengguna ingin mendapatkan lema moden a, langkah pertama ialah semua lema a’ daripada lema a diperolehi menggunakan leksikon moden dan kemudiannya alat padanan digunakan bagi mengumpul semua perkataan a” yang sepadan dari senarai indeks untuk dipadankan kepada beberapa perkataan a’. Dalam proses indeks, leksikon tidak diperlukan. Semasa luar talian, perkataan a” daripada koleksi dokumen mula dipadankan kepada bentuk fleksi moden a’ dan kemudiannya di lema

menggunakan leksikon. Dalam proses pengindeksan, hanya senaraikan lema moden a dengan petunjuk kepada semua bentuk a” yang berlaku(Gotscharek et al. 2011).

Dalam kajian Hyvönen et al. (2011), beliau menyatakan bahawa nama tempat dan liputan geografi berubah mengikut masa. Isu ini menyebabkan masalah ketika mencapai maklumat yang berkaitan dengan masa yang berlainan. Biasanya, kandungan maklumat Geo diindeks menggunakan masa mengenai nama tempat (contohya tombak digunakan dalam peperangan Punic pada tahun 146 di Carthago tetapi diindeks pada masa lain menggunakan nama tempat pada waktu itu). Situasi ini menyebabkan pengguna mempersoalkan kandungan dari segi nama tempat semasa (e.g. Check Republic or Slovakia) atau bertindih nama bersejarah pada waktu berlainan. Beliau mencadangkan pendekatan berdasarkan ontologi untuk menyelesaikan masalah ini. Idea ini adalah untuk mewakili dan mengekalkan siri masa ontologi spatial dari segi perubahan spatial-temporal tempatan yang mudah diurus yang mana ontologi siri masa sebenar boleh dijana secara automatik dengan pengayaan semantik. Ontologi ini kemudiannya boleh digunakan untuk pengindeksan dan pemetaan kawasan spatial-temporal dan nama mereka ke atas satu sama lain. Sebagai bukti-konsep, sistem ini telah digunakan untuk memodelkan perbandaran sejarah Finland pada tahun 1865-2010. Sistem ini telah digunakan dalam portal budaya warisan semantik CultureSampo untuk carian dan cadangan semantik, serta perkhidmatan luaran untuk mengindeks kandungan warisan kebudayaan, dan untuk pengembangan kueri carian dalam sistem pangkalan data warisan budaya.

Abdelli et al. (2015) menjalankan penyelidikan capaian dokumen yang memfokuskan kepada struktur logik yang mewakili komponen seperti bab, bahagian, perenggan, tajuk, tajuk bab dan sebagainya. Tajuk dokumen adalah penting dan mempunyai makna bagi menghasilkan petunjuk kepada sesuatu perenggan. Oleh sebab itu, kajian ini memberi perhatian khusus pada tajuk dokumen semasa proses pengindeksan. Namun, jumlah tajuk dokumen yang terhad ini menyumbang kepada hasil capaian yang tidak relevan. Kajian ini mencadangkan salah satu penyelesaian adalah memanjangkan tajuk dengan menambah istilah lain yang mempunyai kesamaan semantik dengan istilah asal. Eksperimen ini dijalankan pada jumlah korpus

yang besar iaitu INEX 2009 agi menunjukkan keberkesanan cadangan dan penambahbaikan ketepatan keputusan IR.

Järvelin et al. (2016) menjalankan penyelidikan dalam bidang warisan budaya yang memfokuskan kepada koleksi akhbar sejarah. Beliau menyatakan bahawa pendigitan adalah satu cara yang baik untuk memelihara bidang ini dan memudahkan penyelidik dan orang awam mengakses kepada sesuatu maklumat. Oleh itu, aktiviti pendigitan dokumen sejarah sentiasa berkembang. Proses perubahan koleksi cetakan warisan budaya kepada sumber digital yang mudah diakses dan dicari ini adalah dengan menggunakan pengecaman aksara optik (OCR). OCR ini menukarkan dokumen imej terdigit kepada teks digital. OCR boleh mengenalpasti aksara sehingga mencapai ketepatan 99% dari dokumen asal yang mempunyai kualiti imej yang baik. Secara amnya, lebih lama sesuatu akhbar itu, semakin rendah kadar ketepatan yang mungkin didapati. Menurut holley 2009, kadar ketepatan aksara berbeza dari 71% hingga 98% dalam satu sampel pendigitan akhbar dari tahun 1803-1954 yang mana kadar terendah menunjukkan hampir setiap aksara ketiga diakui keliru dan semua perkataan mempunyai kesilapan. Oleh itu, pendigitan koleksi warisan budaya ini sering penuh dengan kesalahan OCR yang menghalang prestasi capaian maklumat. Selain itu, capaian dokumen sejarah ini turut disebut sebagai masalah capaian maklumat silang Bahasa yang mana merujuk kepada capaian dokumen yang ditulis dalam Bahasa sejarah menggunakan kueri moden. Kebanyakan kajian capaian dokumen sejarah memfokuskan kepada penterjemahan kueri yang menghasilkan kueri perkataan berlainan pada masa capaian. Oleh sebab itu, kajian ini menggunakan pendekatan anggaran padanan rentetan dalam mencapai perkataan kueri moden dalam akhbar Finnish terdigit.

Garozzo et al. (2017) mengkaji warisan budaya dengan memberi tumpuan khusus kepada bangunan sejarah yang bercirikan keagamaan. Kajian ini mencadangkan CULTO yang merupakan alat warisan kebudayaan berdasarkan ontologi bagi menyokong pakar warisan kebudayaan di dalam penyiasatan mereka. Alat ini secara khusus boleh menyokong anotasi, pengindeksan automatik, klasifikasi dan pengumpulan data fotografi dan dokumen teks bersejarah bangunan. CULTO turut berfungsi sebagai alat berguna untuk Pemodelan Maklumat Bangunan

Bersejarah (H-BIM) dengan membolehkan data 3D semantic pemodelan dan pengayaan selanjutnya dengan maklumat bukan geometri bangunan bersejarah melalui kemasukan konsep baru mengenai dokumen sejarah, imej, pecahan atau bukti ubah bentuk serta elemen hiasan ke dalam platform BIM. CulTO merupakan hasil usaha penyelidikan bersama antara Laboratorium Ukur dan Fotogrametri Senibina "Luigi Andreozzi "dan Lab PeRCeiVe (Pengiktirafan Corak dan Makmal Visi Komputer) Universiti Catania.

Kesimpulannya, berdasarkan ulasan di atas, kajian dalam bidang sejarah menyatakan bahawa bidang IR adalah penting bagi menyokong bidang ini kerana ia melibatkan capaian, kegunaan dan hubungkait dokumen sejarah untuk berkomunikasi memahami peristiwa masa lalu. Namun keseluruhan teknik tidak mempunyai elemen penting yang telah membawa kepada usulan dalam penyelidikan tesis ini. Elemen penting dalam tesis ini merujuk kepada elemen peristiwa di mana berdasarkan keseluruhan kajian lepas tidak mempunyai elemen peristiwa pada hasil carian. Kebanyakan kajian di atas hanya merangkumi IR dokumen sejarah di dalam aspek geografi, aspek temporal, aspek menjawab soalan dan aspek varian ejaan antara perkataan moden dan sejarah (Corda 2007; Elena et al. 2010; Gotscharek et al. 2011; Jones et al. 2001; Koolen et al. 2006; Mirzaee et al. 2004, 2005; Pilz et al. 2006; Schockaert et al. 2010). Kebanyakan kajian yang terdahulu menunjukkan terdapat banyak peluang untuk diperbaiki. Justeru itu, pembaikan capaian maklumat untuk dokumen sejarah boleh diperluaskan kepada elemen peristiwa seperti mana yang disebut oleh JosephandJanda (2008) iaitu kajian sejarah sering memberi tumpuan kepada peristiwa. Selain itu, hanya beberapa kajian menggunakan pendekatan ontologi sebagai pendekatan capaian(Abdelli et al. 2015; Alonso et al. 2007; Corda 2007; Garozzo et al. 2017; Hyvönen et al. 2011; Mirzaee et al. 2005). Namun dalam semua kajian tersebut hanya beberapa kajian yang menyatakan kaedah pembangunan ontologi seperti Jadual 2.4. Oleh itu, semua kaedah pembangunan ontologi yang dikenalpasti dibincangkan dalam kajian ini bagi membantu pembangunan ontologi baru dalam seksyen 2.8. Kajian tersebut juga tidak memberi tumpuan kepada proses pengindeksan dan pemangkatan. Ringkasan analisis kajian capaian maklumat dokumen sejarah boleh dirujuk pada Jadual 2.4.

Jadual 2.4 Ringkasan analisis kajian

Kajian	Kaedah pembangunan ontologi	Pendekatan capaian	Skop kajian	Hasil kajian
Jone(2001)		Guna kueri nama sesuatu tempat	Geografi	Ontologi tempat
Mirzaee(2004)	Model perusahaan	Ontologi	Temporal	Aplikasi semantik dalam dokumen sejarah
Koolen(2006)	-	Pendekatan bahasa silang untuk capaian dokumen bersejarah	Ejaan	Hasil yang positif dalam menyelesaikan isu yang berkaitan dengan capaian
Pilz(2006)	-	Carian berasaskan peraturan	Ejaan	Enjin carian kabur berasaskan peraturan
Mirzaee(2005), Alonso(2007)	Model perusahaan	Ontologi	Temporal	Dokumen diisih dalam aspek temporal
Corda(2007)	METHONTOLOGY, Gruninger & Fox, Uschold & King, Ordnance Survey	Ontologi	Sistem kecekapan menjawab soalan	Aplikasi semantik dalam dokumen sejarah
Hauser(2007)	-	Pendigitalan dokumen sejarah	Ejaan	Capaian kepada dokumen sejarah tidak berjaya
Schockaert(2010)	-	Teknik heuristik mudah	Temporal	Satu rangka kerja berdasarkan pengaburan daripada Aljabar selang Allen tentang capaian dokumen sejarah

bersambung