VOT 74074

AN INTELLIGENT DATA MAPPING FOR HYDROLOGICAL INFORMATION SISTEM (HIS) USING CUBE DATABASES TO CATER FROM VARIOUS DATA TYPE

(PENGEKSTRAKAN DATA PINTAR UNTUK SISTEM MAKLUMAT HIDROLOGI (SMH) MENGGUNAKAN PANGKALAN DATA KIUB UNTUK PEMETAAN DATA YANG MEMPUNYAI PELBAGAI FORMAT)

HARIHODIN BIN SELAMAT
MOHD SHAFRY BIN MOHD RAHIM
DAUT BIN DAMAN

RESEARCH VOTE NO:
74074

Fakulti Sains Komputer Dan Sistem Maklumat
Universiti Teknologi Malaysia

2005

UNIVERSITI TEKNOLOGI MALAYSIA

BORANG PENGESAHAN
LAPORAN AKHIR PENYELIDIKAN

TAJUK PROJEK : **An Intelligent Data Mapping For Hydrological Information System (HIS) Using Cube Database to Cater from Various Data Types**

Saya _____PROF MADYA DR HARIHODDIN SELAMAT_____
(HURUF BESAR)

Mengaku membenarkan **Laporan Akhir Penyelidikan** ini disimpan di Perpustakaan Universiti Teknologi Malaysia dengan syarat-syarat kegunaan seperti berikut :

1. Tesis adalah hakmilik Universiti Teknologi Malaysia.
2. Perpustakaan Universiti Malaysia dibenarkan membuat salinan untuk tujuan rujukan sahaja.
3. Perpustakaan dibenarkan membuat penjualan salinan Laporan Akhir Penyelidikan ini bagi kategori TIDAK TERHAD.
4. *Sila tandakan (✓)

☐	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
☐	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dibuat)
☑	TIDAK TERHAD	

_____
(TANDATANGAN KETUA PENYELIDIK)

_____
Nama & Cop Ketua Penyelidik

Tarikh:_____

CATATAN: * *Jika Laporan Akhir Penyelidikan ini SULIT atau TERHAD, Sila Lampirkan surat daripada pihak berkuasa/ organisasi berkenaan dengan menyatakan sekali sebab dan tempoh laporan ini perlu dikelaskan sebagai SULIT dan TERHAD*

ABSTRACT

Information Extraction is a process that extracts information from existing system source and stores into a database. Previous researchers had focus on information extraction for HTML data using wrapper approach. The drawback from this approach is resiliency where wrapper fails to function when the file of interest's structure changes. Ontology based information extraction is an alternative solution for this problem. In this research, ontology based information extraction used hydrological data from Jabatan Pengairan dan Saliran (JPS) as the case study. Ontology based information extraction for hydrology domain or also known as 'EkstrakPro' is divided into three main processes; which are ontology parser process, keyword and sequences recognition process, and a data mapping process. 'EkstrakPro' used two inputs; the hydrology data and ontology extraction. An important feature in 'EkstrakPro' is that ontology extraction, where unit object is introduced to simplify the ontology maintenance. The sequential recognition algorithm is to solve the time consuming issues for extracting sequential data. Five types of hydrological data are used in the experiment. These data are divided into three categories; (i) original data taken from gauging machine, (ii) the altered data and (iii) the different sizes of data. Based on these categories, the information extraction resiliency and time taken have been measured using a precise equation and O-notation. The results show that prototype 'EkstrakPro' can extract different structure hydrology data correctly by using only one algorithm. Using sequential recognition algorithm can also further reduce the time required for extraction of information. The result of the research proves that information extraction can be solved using ontology approach.

ABSTRAK

Pengekstrakan maklumat merupakan satu proses yang mengekstrak maklumat daripada sumber sistem sedia ada dan menyimpannya ke dalam pangkalan data. Penyelidikan terdahulu tertumpu kepada pengekstrakan maklumat data HTML menggunakan pendekatan *wrapper*. Kelemahan pendekatan ini adalah dari segi ketahanan di mana *wrapper* gagal berfungsi dengan baik jika terdapat perubahan pada struktur fail yang ingin di ekstrak. Pengekstrakan maklumat berasaskan ontologi merupakan penyelesaian alternatif kepada masalah ketahanan. Di dalam penyelidikan ini, pengekstrakan maklumat berasaskan ontologi menggunakan data hidrologi dari Jabatan Pengairan dan Saliran (JPS) sebagai kajian kes. Pengekstrakan maklumat ontologi bagi domain hidrologi dikenali sebagai ‘EkstrakPro’ terbahagi kepada tiga proses utama; iaitu proses penghuraian ontologi, proses pengecam jujukan dan kata kunci serta proses pemetaan data. ‘EkstrakPro’ menggunakan dua input; data hidrologi dan ontologi pengekstrakan. Ciri penting ‘EkstrakPro’ adalah ontologi pengekstrakan, di mana unit objek diperkenalkan bagi memudahkan selenggara ontologi. Algoritma pengecam jujukan menyelesaikan isu penggunaan masa dalam mengekstrak data berjujukan. Lima jenis data hidrologi digunakan di dalam eksperimen. Data-data ini dibahagikan kepada tiga kategori; (i) Data asal daripada mesin bacaan, (ii) data yang diubahsuai dan (iii) perbezaan saiz data. Berdasarkan kategori tersebut, ketahanan pengekstrakan maklumat dan masa yang digunakan dapat diukur menggunakan rumusan ketepatan dan notasi-O. Keputusan menunjukkan prototaip ‘EkstrakPro’ boleh mengekstrak data hidrologi dengan struktur yang berbeza dengan tepat dan menggunakan hanya satu algoritma. Algoritma pengecam jujukan boleh juga mengurangkan masa yang diperlukan oleh pengekstrakan maklumat. Hasil penyelidikan ini membuktikan masalah pengekstrakan maklumat dapat diselesaikan dengan pendekatan ontologi.

"Kami akui karya ini adalah hasil kerja kami sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya jelaskan sumbernya"

Tandatangan	:_____
Nama Ketua Penyelidik	: PROF MADYA DR HARIHODIN SELAMAT
Tarikh	: 29.12.2005

Tandatangan	:_____
Nama Penyelidik I	: MOHD SHAFRY MOHD RAHIM
Tarikh	: 29.12.2005

Tandatangan	:_____
Nama Penyelidik II	: PROF MADYA DAUT DAMAN
Tarikh	: 29.12.2005

PENGHARGAAN

Syukur ke hadrat Ilahi kerana dengan izinNya laporan ini dapat disiapkan. Setinggi-tinggi penghargaan kepada semua pihak yang terlibat atas bimbingan dan penyeliaan yang diberi sepanjang tempoh penyelidikan dijalankan.

Sekian, terima kasih.

ABSTRAK

Pengekstrakan maklumat merupakan satu proses yang mengekstrak maklumat daripada sumber sistem sedia ada dan menyimpannya ke dalam pangkalan data. Penyelidikan terdahulu tertumpu kepada pengekstrakan maklumat data HTML menggunakan pendekatan *wrapper*. Kelemahan pendekatan ini adalah dari segi ketahanan di mana *wrapper* gagal berfungsi dengan baik jika terdapat perubahan pada struktur fail yang ingin di ekstrak. Pengekstrakan maklumat berasaskan ontologi merupakan penyelesaian alternatif kepada masalah ketahanan. Di dalam penyelidikan ini, pengekstrakan maklumat berasaskan ontologi menggunakan data hidrologi dari Jabatan Pengairan dan Saliran (JPS) sebagai kajian kes. Pengekstrakan maklumat ontologi bagi domain hidrologi dikenali sebagai ‘EkstrakPro’ terbahagi kepada tiga proses utama; iaitu proses penghuraian ontologi, proses pengecam jujukan dan kata kunci serta proses pemetaan data. ‘EkstrakPro’ menggunakan dua input; data hidrologi dan ontologi pengekstrakan. Ciri penting ‘EkstrakPro’ adalah ontologi pengekstrakan, di mana unit objek diperkenalkan bagi memudahkan selenggara ontologi. Algoritma pengecam jujukan menyelesaikan isu penggunaan masa dalam mengekstrak data berjujukan. Lima jenis data hidrologi digunakan di dalam eksperimen. Data-data ini dibahagikan kepada tiga kategori; (i) Data asal daripada mesin bacaan, (ii) data yang diubahsuai dan (iii) perbezaan saiz data. Berdasarkan kategori tersebut, ketahanan pengekstrakan maklumat dan masa yang digunakan dapat diukur menggunakan rumusan ketepatan dan notasi-O. Keputusan menunjukkan prototaip ‘EkstrakPro’ boleh mengekstrak data hidrologi dengan struktur yang berbeza dengan tepat dan menggunakan hanya satu algoritma. Algoritma pengecam jujukan boleh juga mengurangkan masa yang diperlukan oleh pengekstrakan maklumat. Hasil penyelidikan ini membuktikan masalah pengekstrakan maklumat dapat diselesaikan dengan pendekatan ontologi.

ABSTRACT

Information Extraction is a process that extracts information from existing system source and stores into a database. Previous researchers had focus on information extraction for HTML data using wrapper approach. The drawback from this approach is resiliency where wrapper fails to function when the file of interest's structure changes. Ontology based information extraction is an alternative solution for this problem. In this research, ontology based information extraction used hydrological data from Jabatan Pengairan dan Saliran (JPS) as the case study. Ontology based information extraction for hydrology domain or also known as 'EkstrakPro' is divided into three main processes; which are ontology parser process, keyword and sequences recognition process, and a data mapping process. 'EkstrakPro' used two inputs; the hydrology data and ontology extraction. An important feature in 'EkstrakPro' is that ontology extraction, where unit object is introduced to simplify the ontology maintenance. The sequential recognition algorithm is to solve the time consuming issues for extracting sequential data. Five types of hydrological data are used in the experiment. These data are divided into three categories; (i) original data taken from gauging machine, (ii) the altered data and (iii) the different sizes of data. Based on these categories, the information extraction resiliency and time taken have been measured using a precise equation and O-notation. The results show that prototype 'EkstrakPro' can extract different structure hydrology data correctly by using only one algorithm. Using sequential recognition algorithm can also further reduce the time required for extraction of information. The result of the research proves that information extraction can be solved using ontology approach.

KANDUNGAN

SENARAI JADUAL

SENARAI RAJAH

SENARAI SINGKATAN

AI	-	*Artificial Intelligent*
BYU	-	*Brigham Young University*
CSV	-	*Comma Separated Variable*
IE	-	*Information Extraction*
JPS	-	Jabatan Pengairan dan Saliran
MHIS	-	*Malaysian Hydrology Information System*
MIT	-	*Molecule Information Table*
NPL	-	*Natural Language Processing*
SRM	-	*Single Robust Model*
UO	-	Unit Objek

SENARAI LAMPIRAN

VOT 74074

SPATIAL AND NON-SPATIAL DATABASES ENHANCEMENT FOR HYDROLOGICAL INFORMATION SYSTEM (HIS)

(PENGEKSTRAKAN DATA BERASASKAN PENDEKATAN ONTOLOGI : KES DATA JUJUKAN HIDROLOGI)

HARIHODIN SELAMAT
MOHD SHAFRY MOHD RAHIM
DAUT DAMAN

RESEARCH VOTE NO:
74074

Fakulti Sains Komputer Dan Sistem Maklumat
Universiti Teknologi Malaysia

2005

"Kami akui karya ini adalah hasil kerja kami sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya jelaskan sumbernya"

Tandatangan :_____

Nama Ketua Penyelidik: PROF MADYA DAUT DAMAN

Tarikh : 2.2.2002

Tandatangan :_____

Nama Penyelidik I : PROF MADYA DR HARIHODIN SELAMAT

Tarikh : 2.2.2002

Tandatangan :_____

Nama Penyelidik II : MOHD SHAFRY MOHD RAHIM

Tarikh : 2.2.2002

PENGHARGAAN

Syukur ke hadrat Ilahi kerana dengan izinNya laporan ini dapat disiapkan. Setinggi-tinggi penghargaan kepada penyelia laporan, Prof. Madya Dr. Harihodin Selamat, Prof. Madya Daut bin Daman dan En. Mohd Shafry bin Mohd Rahim atas bimbingan dan penyeliaan yang diberi sepanjang tempoh penyediaan laporan. Saya juga terhutang budi diatas kesudian mereka membiayai pengajian sarjana ini.

Penghargaan yang tidak terhingga juga ditujukan buat isteri tercinta Puteri Suhaiza Sulaiman yang banyak memberi pandangan dan kritikan. Tanpa sokongan beliau, laporan ini tidak akan sama seperti yang dibentangkan disini.

Sekian, terima kasih.

ABSTRAK

Pengekstrakan maklumat merupakan satu proses yang mengekstrak maklumat daripada sumber sistem sedia ada dan menyimpannya ke dalam pangkalan data. Penyelidikan terdahulu tertumpu kepada pengekstrakan maklumat data HTML menggunakan pendekatan *wrapper*. Kelemahan pendekatan ini adalah dari segi ketahanan di mana *wrapper* gagal berfungsi dengan baik jika terdapat perubahan pada struktur fail yang ingin di ekstrak. Pengekstrakan maklumat berasaskan ontologi merupakan penyelesaian alternatif kepada masalah ketahanan. Di dalam penyelidikan ini, pengekstrakan maklumat berasaskan ontologi menggunakan data hidrologi dari Jabatan Pengairan dan Saliran (JPS) sebagai kajian kes. Pengekstrakan maklumat ontologi bagi domain hidrologi dikenali sebagai ‘EkstrakPro’ terbahagi kepada tiga proses utama; iaitu proses penghuraian ontologi, proses pengecam jujukan dan kata kunci serta proses pemetaan data. ‘EkstrakPro’ menggunakan dua input; data hidrologi dan ontologi pengekstrakan. Ciri penting ‘EkstrakPro’ adalah ontologi pengekstrakan, di mana unit objek diperkenalkan bagi memudahkan selenggara ontologi. Algoritma pengecam jujukan menyelesaikan isu penggunaan masa dalam mengekstrak data berjujukan. Lima jenis data hidrologi digunakan di dalam eksperimen. Data-data ini dibahagikan kepada tiga kategori; (i) Data asal daripada mesin bacaan, (ii) data yang diubahsuai dan (iii) perbezaan saiz data. Berdasarkan kategori tersebut, ketahanan pengekstrakan maklumat dan masa yang digunakan dapat diukur menggunakan rumusan ketepatan dan notasi-O. Keputusan menunjukkan prototaip ‘EkstrakPro’ boleh mengekstrak data hidrologi dengan struktur yang berbeza dengan tepat dan menggunakan hanya satu algoritma. Algoritma pengecam jujukan boleh juga mengurangkan masa yang diperlukan oleh pengekstrakan maklumat. Hasil penyelidikan ini membuktikan masalah pengekstrakan maklumat dapat diselesaikan dengan pendekatan ontologi.

ABSTRACT

Information Extraction is a process that extracts information from existing system source and stores into a database. Previous researchers had focus on information extraction for HTML data using wrapper approach. The drawback from this approach is resiliency where wrapper fails to function when the file of interest's structure changes. Ontology based information extraction is an alternative solution for this problem. In this research, ontology based information extraction used hydrological data from Jabatan Pengairan dan Saliran (JPS) as the case study. Ontology based information extraction for hydrology domain or also known as 'EkstrakPro' is divided into three main processes; which are ontology parser process, keyword and sequences recognition process, and a data mapping process. 'EkstrakPro' used two inputs; the hydrology data and ontology extraction. An important feature in 'EkstrakPro' is that ontology extraction, where unit object is introduced to simplify the ontology maintenance. The sequential recognition algorithm is to solve the time consuming issues for extracting sequential data. Five types of hydrological data are used in the experiment. These data are divided into three categories; (i) original data taken from gauging machine, (ii) the altered data and (iii) the different sizes of data. Based on these categories, the information extraction resiliency and time taken have been measured using a precise equation and O-notation. The results show that prototype 'EkstrakPro' can extract different structure hydrology data correctly by using only one algorithm. Using sequential recognition algorithm can also further reduce the time required for extraction of information. The result of the research proves that information extraction can be solved using ontology approach.

KANDUNGAN

SENARAI JADUAL

SENARAI RAJAH

SENARAI SINGKATAN

AI	-	*Artificial Intelligent*
BYU	-	*Brigham Young University*
CSV	-	*Comma Separated Variable*
IE	-	*Information Extraction*
JPS	-	Jabatan Pengairan dan Saliran
MHIS	-	*Malaysian Hydrology Information System*
MIT	-	*Molecule Information Table*
NPL	-	*Natural Language Processing*
SRM	-	*Single Robust Model*
UO	-	Unit Objek

SENARAI LAMPIRAN

BAB 1

PENGENALAN

1.1 Pendahuluan

Bidang *Information Extraction* (IE) adalah satu bidang yang melakukan proses pengekstrakan maklumat daripada data digital. Youn (1992) mendefinisikan pengekstrakan maklumat sebagai satu proses untuk mengekstrak maklumat daripada sumber sistem sedia ada dan seterusnya menyimpannya ke dalam satu fail. Manakala Xiaoying dan Mengjie (2004) mendefinisikan IE sebagai satu proses yang mengambil fail teks sebagai input dan menghasilkan data mengikut format yang diperlukan. Data ini mungkin dipaparkan kepada pengguna, disimpan di dalam pangkalan data atau *spreadsheet* bagi kegunaan analisis.

Di antara kepentingan IE yang dikenal pasti adalah membantu enjin pencarian dokumen daripada halaman web. Teknik pengekstrakan diperlukan dalam mencari maklumat yang tepat daripada satu atau lebih dokumen web. Selain itu IE diperlukan dalam proses pemindahan data daripada sistem asal ke sistem yang baru. Situasi ini sering berlaku apabila pengguna bertukar sistem komputer. Data daripada sistem asal akan di ekstrak dan diubah format yang sesuai dengan sistem yang baru.

Terdapat beberapa pendekatan IE termasuklah bahasa pembangunan *wrapper,* penggunaan struktur data, *Natural Language Processing* (NLP), permodelan dan ontologi. Tumpuan kebanyakan penyelidik adalah meningkatkan ketepatan *wrapper* di samping mengurangkan penglibatan pengguna dalam proses pengekstrakan iaitu secara automatik. Kelemahan utama sistem IE yang

menggunakan pendekatan *wrapper* adalah ia hanya dapat mengekstrak maklumat daripada data dalam berformat yang terhad dan tertentu sahaja.

Sementara itu, terdapat sekumpulan penyelidik daripada Universiti Brigham Young sedang berusaha meningkatkan penggunaan konsep skema yang lebih umum bagi meningkatkan ketepatan IE. Kumpulan ini mula memperkenalkan pendekatan ontologi di dalam IE (Embley et al., 1998). Ontologi adalah spesifikasi dalam membentuk suatu konsep (Gruber, 1993). Dari sudut bidang falsafah, ontologi merujuk kepada suatu kewujudan. Di dalam konsep perkongsian pengetahuan (*knowledge sharing*) aplikasi kepintaran buatan (AI), ontologi adalah penerangan mengenai konsep dan hubungan yang wujud bagi satu agen. Kelebihan utama IE berasaskan ontologi adalah mempunyai ketahanan pengekstrakan maklumat. Menyedari kelebihan ini, bidang IE berasaskan ontologi akan menjadi fokus penyelidikan ini.

1.2 Latar Belakang Masalah

Penggunaan data digital telah berkembang pesat beberapa tahun kebelakangan ini. Ini kerana dorongan penggunaan *world web wide* (www) yang semakin meningkat. IE digunakan bagi mengekstrak maklumat daripada fail HTML. Pendekatan seperti bahasa *wrapper* (Crescenzi et al., 2001; Hammer et al., 1997; Arocena dan Mendelzon, 1998), NLP (Calif dan Mooney, 1999; Freitag, 2000;Sonderlan,1999) dan permodelan (Adelberg, 1998) diperkenalkan bagi mengekstrak maklumat yang diperlukan pengguna. Walaupun kebanyakan penyelidik melaporkan kejayaan hasil daripada pengujian yang dilakukan, namun pendekatan ini masih mempunyai masalah ketahanan. Kelemahan dari segi ketahanan bermakna sebuah *wrapper* akan gagal berfungsi dengan baik sekiranya terdapat perubahan pada struktur fail yang ingin di ekstrak.

IE berasaskan ontologi adalah penyelesaian kepada masalah ketahanan. Pengekstrakan maklumat ontologi adalah model konsepsi yang menerangkan aplikasi

dunia sebenar dengan terperinci. Ciri penting pendekatan ini adalah ontologi pengekstrakan yang dihasilkan daripada data dalam sesebuah bidang tanpa bergantung kepada struktur fail input.

Oleh sebab kebanyakan IE berasaskan ontologi hanya tertumpu kepada fail HTML, timbul persoalan, apakah pendekatan ini boleh digunakan ke atas dokumen lain selain fail HTML? Dalam penyelidikan kali, kajian akan dilaksanakan ke atas IE berasaskan ontologi dengan menggunakan fail teks. Ini kerana fail teks mengandungi sedikit penunjuk untuk mengenal pasti struktur berbanding dengan fail HTML. Fail HTML mempunyai penunjuk-penunjuk yang membezakan struktur antara permulaan <head>, tajuk <title>, kandungan <body> dan sebagainya. Sementara itu tidak semua elemen di dalam fail teks dipisahkan dengan tanda atau tag HTML. Maka proses IE daripada fail teks adalah lebih sukar daripada fail HTML (Adelberg, 1998).

Menyedari kekurangan penyelidikan ke atas IE berasaskan ontologi bagi data selain HTML, penyelidikan ini telah memilih untuk mengkaji keberkesanan IE berasaskan ontologi dalam mengekstrak data hidrologi. Satu kajian kes dilakukan ke atas *Malaysian Hidrology Information System* (MHIS) dari Jabatan Pengairan dan Saliran (JPS), yang mana sebelum ini menggunakan pendekatan pengekstrakan data secara tradisional. Penerangan dan kelemahan MHIS akan dibincangkan pada Bahagian Kajian Kes.

1.3 Kajian Kes

MHIS di Jabatan Pengairan dan Saliran (JPS) telah dibangunkan dengan usaha sama Universiti Teknologi Malaysia (UTM) dan *Water Institute*, UK. MHIS digunakan untuk menyimpan dan manipulasi maklumat hidrologi yang terdiri daripada beberapa modul antaranya adalah perisian *MHIS Dataload*. Modul ini menyediakan kemudahan untuk memindahkan data hidrologi ke dalam sistem pangkalan data MHIS (Jabatan Pengairan dan Saliran, 2001a).

MHIS Dataload terdiri daripada beberapa algoritma yang dibangunkan khas bagi data taburan hujan, penyejatan, aras air sungai, enapan terapung dan kualiti air. Algoritma pengekstrakan data telah ditulis di dalam atur cara secara tetap (*hardcoded*) bagi setiap jenis data-data di atas. Proses penyenggaraan perisian ini memerlukan banyak usaha dan masa. Berikut adalah beberapa kelemahan *MHIS Dataload* yang telah dikenal pasti :

1. Algoritma mengekstrak data tidak dinamik. Maka algoritma perlu dikemas kini apabila perubahan struktur atau format data berlaku. Perisian perlu dikemaskinikan setiap kali berlaku perubahan struktur data.

2. Satu algoritma digunakan bagi satu jenis data hidrologi. Maka apabila satu jenis data hidrologi baru digunakan, ia memerlukan satu algoritma pengekstrakan yang baru.

3. Algoritma bergantung kepada struktur dan format data. Data yang dihasilkan oleh manusia selalunya mempunyai banyak ralat atau kesilapan. Data yang akan di ekstrak perlu dibersihkan daripada kesilapan dan ralat.

Berdasarkan kelemahan-kelemahan di atas, persoalan yang dikaji adalah apakah IE berasaskan ontologi sesuai untuk data hidrologi dan sekali gus dapat mengatasi kelemahan-kelemahan yang dihadapi oleh MHIS *Dataload* ?

1.4 Motivasi Kajian Kes

Penyelesaian yang dihasilkan di dalam penyelidikan ini akan dapat membantu dalam mempertingkatkan kecekapan dan ketepatan kerja-kerja pemindahan data hidrologi di dalam bentuk teks ke dalam pangkalan data *MHIS* di JPS.

1.5 Pernyataan Masalah Penyelidikan

Tujuan penyelidikan ini adalah untuk mengkaji IE berasaskan ontologi dengan menggunakan fail teks hidrologi JPS. Dengan implementasi ontologi pengekstrakan ke atas bidang data hidrologi, perkara berikut perlu diperjelaskan.

1. Bagaimana menghasilkan ontologi pengekstrakan bagi mencapai matlamat penyelidikan?
2. Bagaimana menyatakan dengan cara teratur bagi setiap kata kunci, prosa bidang data hidrologi?
3. Bagaimana maklumat diasingkan daripada sumber data berdasarkan kata kunci di dalam ontologi?
4. Bagaimana menentukan keberkesanan IE berasaskan ontologi mengekstrak maklumat daripada fail teks hidrologi.
5. Apakah pembaikan yang boleh dilakukan ke atas IE berasaskan ontologi dalam mengekstrak fail teks hidrologi.

1.6 Matlamat Penyelidikan

Mengkaji keberkesanan IE berasaskan ontologi dalam mengekstrak maklumat daripada fail teks bidang hidrologi.

1.7 Objektif Penyelidikan

Objektif penyelidikan adalah seperti berikut :

1. Membina ontologi pengekstrakan bagi menterjemahkan kata kunci dan hubungan kata kunci fail teks hidrologi.
2. Membina algoritma pengecam jujukan bagi mengurangkan masa pengekstrakan.
3. Melakukan pengujian pengekstrakan maklumat daripada fail teks hidrologi.

1.8 Skop Penyelidikan

1. Fail yang digunakan adalah fail teks berjujukan, yang mana bentuk jujukan adalah konsisten. Fail input yang digunakan adalah data hidrologi daripada JPS, yang mana ia berada di dalam bentuk berjujukan.
2. Struktur pangkalan data yang digunakan berdasarkan skema yang dijana daripada ontologi pengekstrakan.
3. Ontologi pengekstrakan dihasilkan secara manual bagi menghasilkan ekspresi yang lengkap agar matlamat penyelidikan dicapai.

1.9 Sumbangan Ilmiah

Sumbangan akhir penyelidikan adalah seperti berikut :

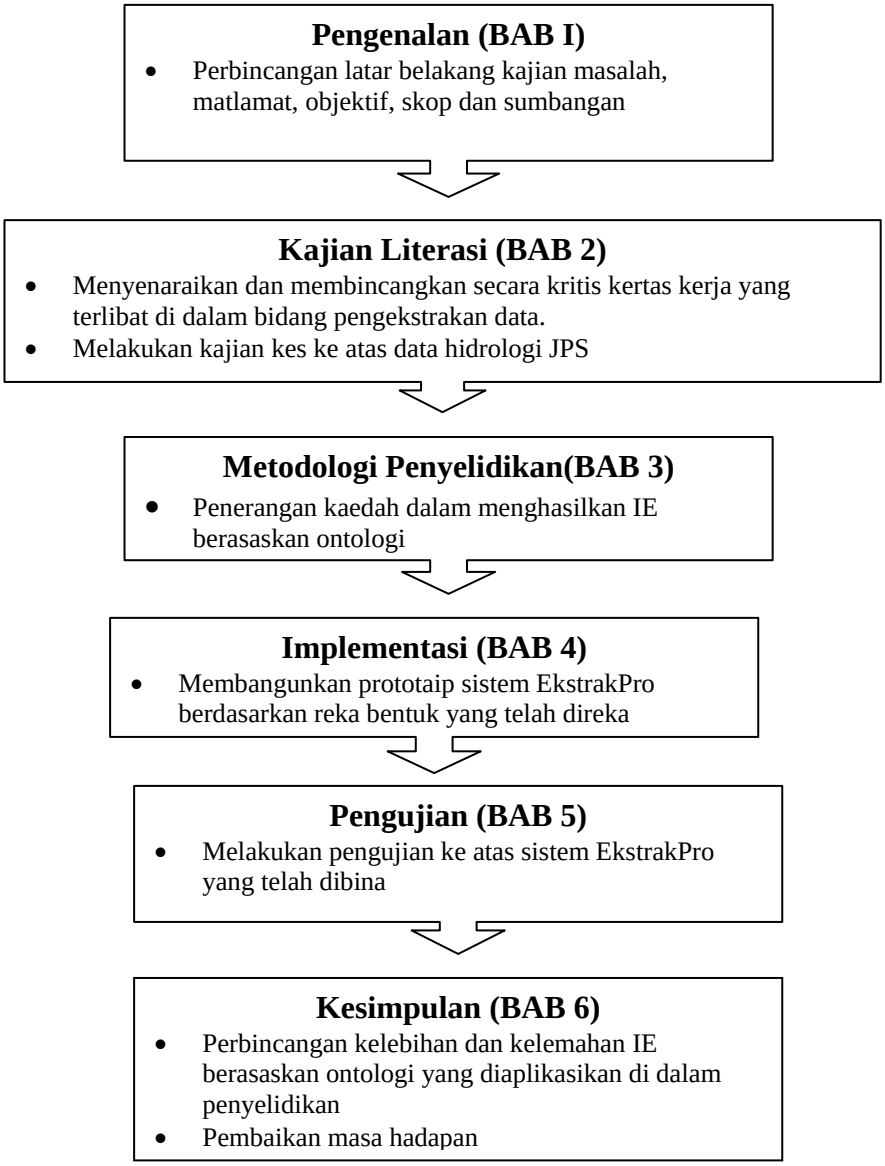
1. Di dalam penyelidikan ini, IE berasaskan ontologi telah digunakan ke atas bidang data hidrologi. Kajian menunjukkan bahawa IE berasaskan ontologi dapat digunakan ke atas fail teks berjujukan.

2. Unit Objek (UO) diperkenalkan bagi menyatakan corak kata kunci. UO adalah kaedah menghasilkan kata kunci secara sistematik. Penggunaan UO dapat mengurangkan kesilapan di dalam menghasilkan kata kunci.

3. Penghasilan algoritma jujukan dalam meningkatkan kepantasan masa proses pengekstrakan bagi data berjujukan. Algoritma pengecam jujukan berfungsi sebagai pembaca bentuk jujukan maklumat. Jika bentuk jujukan telah dikenal pasti, maklumat akan di ekstrak tanpa membandingkan kata kunci dan fail teks. Dengan ini dapat masa proses pengekstrakan dapat dipercepatkan.

1.10 Struktur Laporan

Laporan ini secara keseluruhannya terbahagi kepada 6 bab seperti ditunjukkan di dalam Rajah 1.1.



Rajah 1.1 : Struktur Laporan

BAB 2

KAJIAN LITERASI

2.1 Pendahuluan

Bab ini akan membincangkan beberapa kategori pengekstrakan maklumat (IE) seperti pendekatan bahasa, HTML, induksi, model, NPL dan ontologi. Pemilihan pendekatan yang sesuai dilakukan berdasarkan kajian kes dengan mengambil kira format data hidrologi yang digunakan. Seterusnya perbincangan dilakukan ke atas beberapa metodologi di dalam pembinaan ontologi pengekstrakan hasil penyelidikan terdahulu bagi memilih metodologi yang terbaik. Selain itu, bab ini turut memberi penerangan bagi setiap jenis data hidrologi JPS yang digunakan di dalam penyelidikan ini.

2.2 Pengekstrakan Maklumat (IE)

Penyelidikan ke atas IE banyak tertumpu kepada halaman web. Untuk mengekstrak maklumat daripada halaman web, satu agen perlu mengesan maklumat yang dikehendaki. Percubaan yang terawal dalam mengekstrak maklumat daripada web secara automasi melibatkan penghasilan *wrapper* bagi halaman yang dikehendaki secara manual. *Wrapper* yang dihasilkan adalah khusus untuk halaman yang tertentu mengakibatkan kelemahan dalam proses penghasilannya yang remeh, dan perlu dihasilkan semula sekiranya halaman berkenaan berubah. Oleh kerana ini, ramai penyelidik tertumpu kepada penghasilan *wrapper* secara semi-automasi.

Sehingga kini, terdapat hampir 39 *wrapper* sebagaimana yang dinyatakan oleh Kuhlin (2002). *Wrapper* ini boleh dibahagikan kepada enam kategori iaitu pendekatan bahasa, pendekatan HTML, pendekatan induksi, pendekatan model, pendekatan NPL dan pendekatan ontologi.

- **Bahasa Pembangunan Wrapper**

Salah satu pendekatan terawal yang digunakan untuk menghasilkan penjana *wrapper* adalah pembangunan bahasa yang direka khusus untuk membantu pengguna menghasilkan *wrapper*. Bahasa ini digunakan sebagai alternatif kepada bahasa umum seperti Java dan Perl. Beberapa pengekstrakan yang menggunakan teknik ini adalah Minerva (Crescenzi dan Mecca, 1998), TIMMIS (Hammer et.al, 1997) dan Web-OQL (Arocena dan Mendelzon, 1998).

- **Pendekatan HTML**

Pendekatan pengekstrakan ini bergantung kepada ciri struktur yang diwarisi daripada dokumen HTML untuk melakukan pengekstrakan maklumat. Ia mengesan data berdasarkan lokasi yang telah dihasilkan daripada pohon huraian. Pohon ini adalah perwakilan tag HTML secara hierarki. Pengekstrakan dapat dilakukan secara semi-automatik sekiranya diberi satu contoh, dan automatik sekiranya diberi banyak contoh halaman daripada satu sumber. Antara pengekstrak yang menggunakan pendekatan ini adalah W4F (Sahuguet and Azavant, 2001) dan RoadRunner (Mecca et al., 1998).

- **Pendekatan Induksi**

Pengekstrakan induksi mengenal pasti corak yang terdapat di dalam satu set halaman latihan yang telah dilabel. Perbezaan utama pengekstrakan induksi dengan pengekstrakan yang berasaskan NPL adalah induksi tidak bergantung kepada kekangan linguistik. Malah ianya bergantung kepada format struktur yang

akan menekankan struktur di mana akan ditemui. Ini membuatkan pengekstrakan berasaskan induksi lebih sesuai untuk halaman HTML berbanding teknik sebelum ini. Di antara pengekstrakan yang terdapat di pasaran adalah WIEN (Kushmerick, 2000), SoftMealy (Hsu and Dung, 1998) dan STALKER (Muslea et al., 2001).

- **Pendekatan Model**

Pendekatan ini menggunakan kaedah yang hampir serupa seperti kaedah pendekatan induksi untuk memadankan struktur data yang diberikan oleh pengguna. Pengekstrakan yang menggunakan pendekatan ini adalah NoDoSE (Adelberg, 1998).

- **Pendekatan NPL**

Natural Language Processing (NPL) merupakan satu pendekatan yang digunakan dalam pengekstrakan untuk belajar peraturan pengekstrakan yang dapat mengekstrak maklumat yang dikehendaki dalam dokumen bebas. Pengekstrakan yang berasaskan NPL ini adalah bersesuaian untuk halaman web yang mengandungi teks bertatabahasa, stail telegrafi seperti senarai pekerjaan, iklan sewa rumah, pengumuman seminar dan sebagainya. Antara pengekstrakan yang menggunakan pendekatan ini adalah RAPIER (Calif and Mooney, 1999), SRV (Freitag, 2000) dan WHISK (Sonderlan,1999).

- **Pendekatan ontologi**

Pendekatan ontologi bergantung kepada model konseptual data yang ingin di ekstrak. Ontologi yang diperkenalkan oleh BYU di Universiti Brigham Yoong merupakan perintis bagi kaedah ini (Embley et al., 1998). Kelebihan utama pendekatan ini adalah ianya mudah disesuaikan dalam pelbagai situasi dan juga tahan lasak.

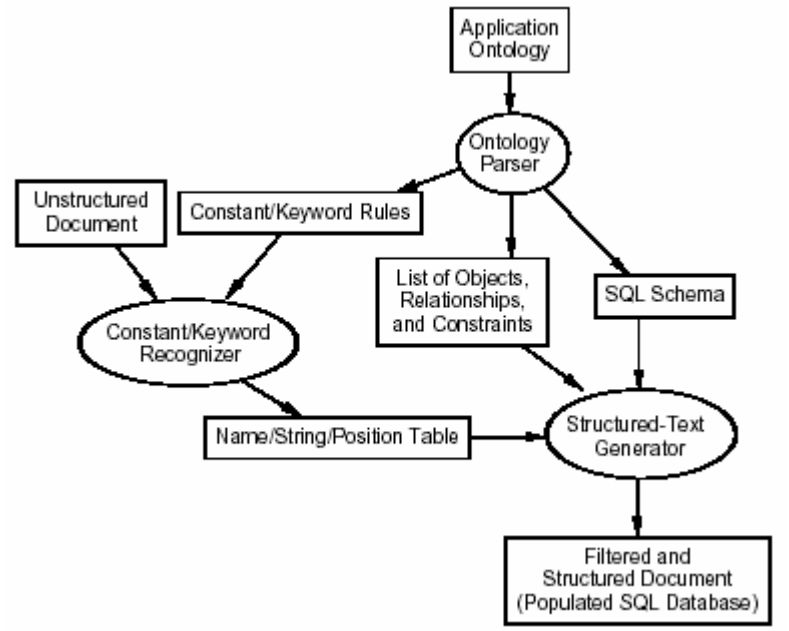
Wrapper yang dihasilkan dari tiga pendekatan pertama (HTML, Induksi dan Model) yang tersenarai di atas hanya dapat mengekstrak data dari halaman serupa dengan halaman latihan. Ia hanya dapat dilaksanakan pada halaman yang sama dari segi formatnya. Ia bermakna latihan perlu dilakukan pada setiap sumber data. Oleh kerana matlamat kajian adalah mengekstrak data hidrologi daripada format data input yang berbeza-beza, maka pendekatan adalah merugikan. Tambahan pula, sekiranya terdapatnya data hidrologi yang baru (dari segi formatnya), maka latihan baru terpaksa dilaksanakan. Pendekatan-pendekatan ini memberi masalah ketika proses penghasilan dan proses penyenggaraannya.

Pendekatan NPL turut tidak sesuai untuk digunakan ke atas data hidrologi kajian kes. Ini kerana pendekatan NPL menggunakan tanda dari struktur ayat yang telah dihuraikan untuk mengenal pasti data yang diperlukan. Data hidrologi tidak mempunyai ayat yang lengkap.

Pendekatan berasaskan ontologi pula amat fleksibel. Ini kerana, ia dapat mengekstrak data tanpa memerlukan set latihan bagi format-format yang berbeza. Berbeza dengan penggunaan *wrapper* di mana setiap format yang berkaitan ke atas sesebuah domain aplikasi, ontologi digunakan bagi merangkumi kesemua data yang berbeza format ke atas satu domain. Oleh kerana ontologi menerangkan domain sebuah subjek berbanding sebuah dokumen. IE berasaskan ontologi adalah tegar ke atas perubahan format data dan dapat mengendalikan data daripada pelbagai sumber tanpa mengganggu gugat ketepatan pengekstrakan. Namun begitu, pendekatan ontologi kurang digunakan berbanding pendekatan *wrapper* yang lain adalah kerana proses penghasilannya yang memerlukan usaha lebih. Pendekatan ontologi adalah pendekatan yang paling sesuai untuk kajian kes kerana menangani pelbagai jenis struktur data teks selain daripada data berstruktur HTML.

2.3 Pengekstrakan berasaskan Ontologi

Rangka kerja sistem yang mengekstrak maklumat struktur daripada dokumen tidak berstruktur berasaskan ontologi mula diperkenalkan oleh Embley et.al (1998). Proses di dalam rangka kerja ini tidak memerlukan campur tangan pengguna dan beroperasi secara automatik. Akan tetapi proses menghasilkan ontologi dibuat secara manual. Dengan menggunakan fail HTML, ujian yang dilaksanakan menunjukkan keputusan dengan ketepatan 99.999%. Ralat di dalam pengujian hanya disebabkan ontologi yang tidak lengkap.



Rajah 2.1: Rangka kerja Pengekstrakan Maklumat berasaskan Ontologi

Rajah 2.1 di atas adalah rangka kerja yang dihasilkan oleh Embley et.al (1998) bagi mengekstrak maklumat daripada data tidak berstruktur. Di dalam rajah tersebut, objek berbentuk kotak mewakili fail manakala bentuk ovul mewakili proses. Input bagi rangka kerja ini adalah ontologi pengekstrakan (*application ontology*) dan dokumen tidak berstruktur (*unstructured document*), dan outputnya adalah dokumen berstruktur (*structured document*). Terdapat tiga proses iaitu penghuraian ontologi (*ontology parser*), pengecam kata kunci (*constant/keyword recognizer*) dan penjana teks berstruktur (*structured text generator*).

Ontologi pengekstrakan adalah ekspresi setiap konteks kata kunci bagi domain yang dikehendaki. Model data *Object-oriented System Model* (OSM) digunakan bagi membina ontologi pengekstrakan. Proses pertama adalah menghurai ontologi pengekstrakan yang akan menghasilkan skema SQL sebagai *createtable statement*. Nama set objek daripada ontologi mewakili atribut jadual SQL yang dijana. Jenis data (*datatype*) *varchar* digunakan bagi setiap atribut bentuk leksikal manakala jenis *integer* bagi objek bukan leksikal. Maklumat hubungan antara objek digunakan dalam deklarasi dan kekangan kardinal skema SQL yang menentukan setiap hubungan sama ada *one to many*, *many to many* dan sebagainya. Akhir sekali penghuraian menghasilkan set peraturan kata kunci.

Proses kedua seterusnya adalah pengecam kata kunci yang menerima input peraturan set kata kunci dan dokumen tidak berstruktur seperti Rajah 2.2. Pengecam mengguna setiap ekspresi kata kunci bagi membandingkan setiap perkataan di dalam dokumen tidak berstruktur. Apabila pengecaman suatu perkataan *S* berdasarkan ekspresi *E* dengan nama *T*. *T* akan dianggap sebagai nama dan *S* sebagai perkataan, set ini dikenali sebagai jadual struktur data. Proses perbandingan setiap perkataan akan menggunakan masa yang banyak jika terdapat rekod sehingga 1000 baris seperti data hidrologi JPS. Sedikit pembaikan perlu dibuat agar masa proses perbandingan kata kunci menggunakan data hidrologi tidak terlalu lama.

```
'96 CHEV Monte Carlo Z34, loaded, bright Red
15,000 actual miles!  A great buy at $14,990,
$750 to 1000 down.  MURDOCK CHEVROLET 298-8090

#####

'94 CHEV Corsica, 88,281 miles.  Ask for #16. $4,900.
Government Surplus533-5885

#####

'89 AUDI 80, red, auto., p/w, p/l, sunroof, loaded, 128K,
new trans., new diff.  Runs perfect, must sell, $3300 obo.
gcall Nate, 554-4414
```

Rajah 2.2: Contoh Dokumen tidak Berstruktur

Bagi proses ketiga, penjanaan teks berstruktur dilaksanakan menggunakan input skema SQL dan senarai objek/hubungan dan kekangan bagi memadankan objek dengan jadual struktur data. Pemadanan dilaksanakan secara heuristik:

- Persamaan kata kunci

  Jika kekangan dalam ontologi memerlukan sekurang-kurang satu *constan* bagi satu set objek, dan jika terdapat konteks kata kunci bagi set objek di dalam jadual struktur data, sistem akan menyingkirkan semua *constan* kecuali ia sama nama dengan nama set objek.

- Mengumpul dan Pertindihan *constan*

  Pengecam kata kunci akan menggabungkan perkataan tunggal di dalam sumber dengan lebih nama set objek. Tetapi bagi perkataan yang diberi daripada teks mungkin hanya menghasilkan *constan* tunggal. Oleh itu jika terdapat pertindihan *constan*, sistem akan menyingkir semua kecuali satu *constan*. Penyingkiran *constan* bermula dengan *constan* yang tidak berkait dengan kata kunci.

- Fungsi hubungan

  Jika ontologi menetapkan pangkalan data boleh menerima banyak *constan* bagi satu objek, O dan terdapat satu *constan* bagi O, simpan *constan* ke dalam pangkalan data.

- Bukan fungsi hubungan

  Jika ontologi menetapkan pangkalan data boleh menerima banyak *constan* bagi satu objek dan jika terdapat satu atau lebih *constan*, simpan semua ke dalam pangkalan data.

- Pertama kali tanpa kekangan

Jika ontologi menetapkan pangkalan data boleh menerima sekurang-kurangnya satu *constan* bagi satu objek, O, tetapi jika terdapat beberapa *constan*, simpan *constan* pertama daripada senarai.

2.4 Ontologi Pengekstrakan

Walaupun terdapat ontologi berskala besar, pengkaji ontologi masih perlu membina ontologi bagi domain tertentu, di samping melakukan pengemaskinian terhadap ontologi berkenaan. Pembinaan ontologi secara manual merupakan satu proses memakan masa dan tenaga yang membebankan. Tambahan pula, proses pengemaskinian yang kadang kala dilakukan dengan kadar perlahan, akan menyebabkan masalah terhadap perkembangan aplikasi ontologi berkenaan.

Permulaan bagi penghasilan ontologi berasal dari situasi yang berbeza. Sesebuah ontologi mungkin dibina dari asas, atau sambungan ke atas ontologi sedia ada, atau dari satu sumber informasi bertulis ataupun gabungan kedua-duanya sekali. Pembinaan ontologi turut bervariasi mengikut tahap pengautomasian, antaranya adalah secara manual sepenuhnya, semi-automatik sehingga automatik sepenuhnya. Namun sehingga kini, penghasilan ontologi secara automatik sepenuhnya hanya berkesan ke atas ontologi yang mudah dengan syarat-syarat yang terhad.

Lazimnya, kaedah untuk membina ontologi dapat diringkaskan sebagai : bawah ke atas iaitu dari pengkhususan ke penyeluruhan, atau atas ke bawah iaitu dari penyeluruhan ke pengkhususan; dan tengah keluar (*middle-out*) iaitu dari konsep-konsep penting ke penyeluruhan dan pengkhususan sebagai contoh O*ntologi Enterprise* dan metodologi ontologi yang dicadangkan oleh Lopez (1999). Terdapat beberapa reka bentuk ontologi yang telah dicadangkan oleh beberapa pengkaji ontologi, di antaranya adalah seperti berikut:

- Guarino (1998) memperkenalkan satu metodologi dalam reka bentuk ontologi yang diinspirasikan dari penyelidikan fisiologi yang dikenali

sebagai '*Formal Ontology*' oleh Cocchiarella (1991). Reka bentuk ini mengandungi teori ke atas keseluruhan, teori ke atas bahagian, teori ke atas identiti, teori ke atas kebergantungan, dan teori ke atas universal. Beliau meringkaskan reka bentuk asas perlulah merangkumi :

1. Jelas mengenai domain
2. Menitik beratkan identiti
3. Mengasingkan struktur taksonomi asas
4. Mengenal pasti peranan dengan tepat

- Uschold dan Gruninger (1996) pula memperkenalkan satu rangka metodologi untuk pembinaan ontologi yang dilakukan secara manual sepenuhnya.
    1. Kenal pasti tujuan dan skop
    2. Bina ontologi dalam tiga langkah mudah iaitu
        - Pengenalpastian Ontologi (*Ontology capture*)
          Pengenalpastian konsep asas dan hubungan serta usaha menyediakan definisi bagi objek dan hubungannya.
        - Pengekodan Ontologi (*Ontology coding*)
          Melakukan terma asas untuk ontologi seperti kelas, entiti dan hubungan; memilih bahasa perwakilan dan seterusnya melakukan pengekodan.
        - Pengintegrasian ke atas ontologi sedia ada
    3. Penilaian dan tafsiran
    4. Dokumentasi
    5. Garis panduan untuk setiap langkah sebelumnya

- *Ontological Design Pattern* (ODP) oleh Reich(1999) digunakan untuk mengabstrakkan dan pengenalpastian struktur reka bentuk ontologi, terma, ekspresi dan konteks semantik. Teknik ini dapat dibahagi kepada pembinaan dan pendefinisian ekspresi yang kompleks dari perwakilan asasnya kepada perubahan secara ketidakbergantungan. Teknik ini telah dibuktikan berkesan ketika di aplikasi ke atas informasi molekular biologi.

Hwang (1999) mencadangkan beberapa kriteria yang perlu ada pada sesebuah ontologi yang telah dibina iaitu :

1. Terbuka dan dinamik – baik dari segi algoritma ataupun strukturnya bagi memudahkan pembinaan dan mengemasi.
2. Dapat diukur
3. mudah untuk dikemaskinikan
4. ketidakbergantungan konteks

Daripada teknik yang dibincangkan di atas, teknik oleh Uschold dan Gruninger (1996) adalah sesuai bagi penyelidikan ini. Ini kerana matlamat penyelidikan yang lebih menjurus kepada mengkaji keberkesanan pengekstrakan maklumat ke atas data teks hidrologi. Penghasilan ontologi secara manual akan mengurangkan risiko kegagalan dalam mencapai matlamat tersebut di mana ekspresi ontologi dapat dinyatakan dengan lengkap. Selain itu, ontologi secara manual masih memenuhi objektif pertama penyelidikan.

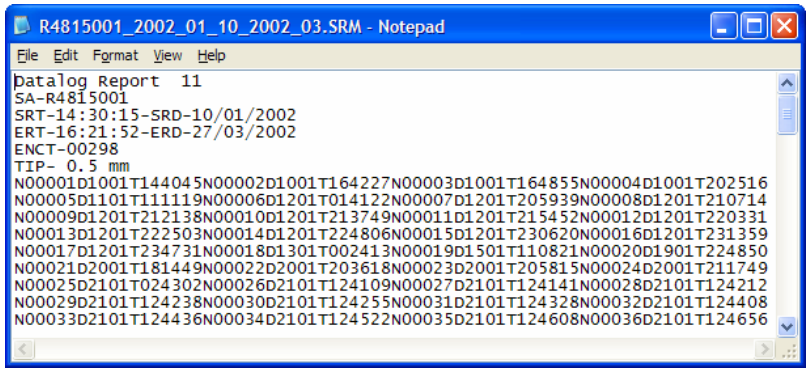
2.5 Kajian Kes ke atas Data Hidrologi JPS

JPS memperoleh data hidrologi daripada 2405 stesen cerapan yang dipasang di seluruh Malaysia. Terdapat 5 jenis data hidrologi yang digunakan di dalam MHIS iaitu data taburan hujan, penyejatan, aras air sungai, enapan terapung sungai dan kualiti air sungai. Maklumat yang ingin di ekstrak daripada setiap data-data ini adalah maklumat id stesen, tarikh, masa catatan serta nilainya bacaannya. Format data hidrologi pula adalah berbeza mengikut jenis cerapannya. Secara asasnya terdapat tiga format yang sedang digunakan bagi penyelidikan iaitu SRM, MIT dan CSV.

2.5.1 SRM

Data format SRM digunakan di dalam menyimpan data perakam taburan hujan elektronik model RF14. Menggunakan kad ingatan (kad SRM) sebagai storan,

format ini akan dibaca oleh program yang dinamakan *hydro reader* dan menyimpan data tersebut ke dalam fail SRM seperti ditunjukkan di dalam Rajah 2.3.



Rajah 2.3 : Contoh keratan format SRM

Sebagaimana yang ditunjukkan dalam rajah 2.1, data yang disimpan di dalam format SRM mengandungi kepala (*header*) dan badan (*body*). Kepala mengandungi enam baris yang akan menyimpan maklumat seperti berikut :

Baris pertama :	Nama fail
Baris kedua :	Nombor stesen
Baris ketiga :	Masa dan tarikh bermula (data diambil)
Baris keempat :	Masa dan tarikh berakhir (data berhenti diambil)
Baris kelima :	Jumlah bilangan data
Baris keenam :	Nilai (iaitu pertambahan nilai untuk setiap masa yang diambil)

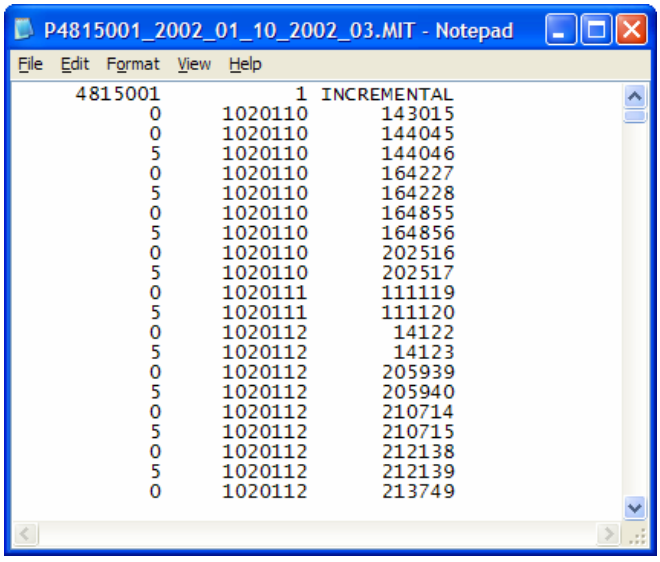
Untuk bahagian badan pula, tata susunannya adalah berterusan tanpa tab atau pun koma. Perbezaan data diwakili daripada huruf pertama sebelum angka iaitu N (bilangan data ke-n), D (tarikh) dan T (masa). Sebagai contoh, keratan data SRM di Rajah 2.3 akan menghasilkan jujukan maklumat seperti yang ditunjukkan dalam Rajah 2.4 di bawah.

N00001D1001***T144045*****N00002**D1001***T164227*****N00003**D1001***T164855***
N00001 D1001 T144045
N00002 D1001 T164227
N00003 D1001 T164855

Rajah 2.4 : Penyusunan format SRM

2.5.2 MIT

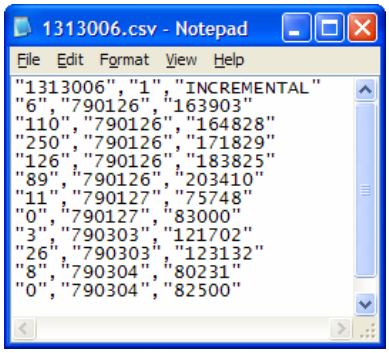
Data yang dicerap menggunakan perakam carta perlu ditukar kepada format digital dengan menggunakan program *Box Car*. *Box Car* akan menghasilkan data di dalam fail MIT, iaitu satu format yang boleh dibaca oleh komputer peribadi (PC) atau UNIX. Penyusunan data MIT adalah lebih mudah berbanding SRM. Ini kerana data MIT telah dibahagikan mengikut lajur-lajur yang terdiri daripada nilai, tarikh dan masa yang dijarakkan melalui tab. Rajah 2.5 berikut adalah contoh keratan data format MIT.



Rajah 2.5 : Contoh keratan format MIT

2.5.3 CSV

Comma-delimited format atau CSV adalah data format data yang umum diguna pakai oleh kebanyakan aplikasi komputer. Format ini digunakan bagi menyimpan data yang dikutip secara manual atau elektronik oleh juruteknik JPS. Dengan menggunakan borang-borang yang disediakan oleh pihak JPS (JPS6 Pin. 3/83, JPT IIB – Pin 2/83, JPT 11C – Pin 1/2000), Maklumat dianalisis dan disimpan di dalam bentuk CSV. Data CSV mengandungi tiga lajur iaitu nilai, tarikh dan masa. Berbeza dengan data MIT, data CSV menggunakan (“) pada mula dan (”) pada akhir maklumat. Setiap lajur dibezakan dengan tanda koma (,). Rajah 2.6 menunjukkan contoh data format CSV.



Rajah 2.6 : Contoh keratan format CSV

2.6 Kesimpulan

Secara keseluruhannya, dapat disimpulkan IE berasaskan ontologi adalah paling sesuai untuk masalah kajian kes yang telah dibincangkan di dalam Bab 1. Ini kerana format data hidrologi itu sendiri yang berbeza dengan format-format data penyelidikan terdahulu iaitu data teks berjujukan sebagaimana yang telah dibincangkan dalam bahagian 2.5. Pemilihan metodologi bagi IE berasaskan ontologi dan ontologi pengekstrakan adalah berdasarkan garis panduan yang dicadangkan oleh Embley et al.(1998) dan Ushold dan Gruininger (1996) kerana pengekstrakan ontologi dilakukan secara manual sepenuhnya. Ini bersesuaian dengan matlamat penyelidikan yang lebih menjurus kepada mengkaji keberkesanan pengekstrakan maklumat berasaskan ontologi ke atas data teks hidrologi.

BAB 3

METODOLOGI PENYELIDIKAN

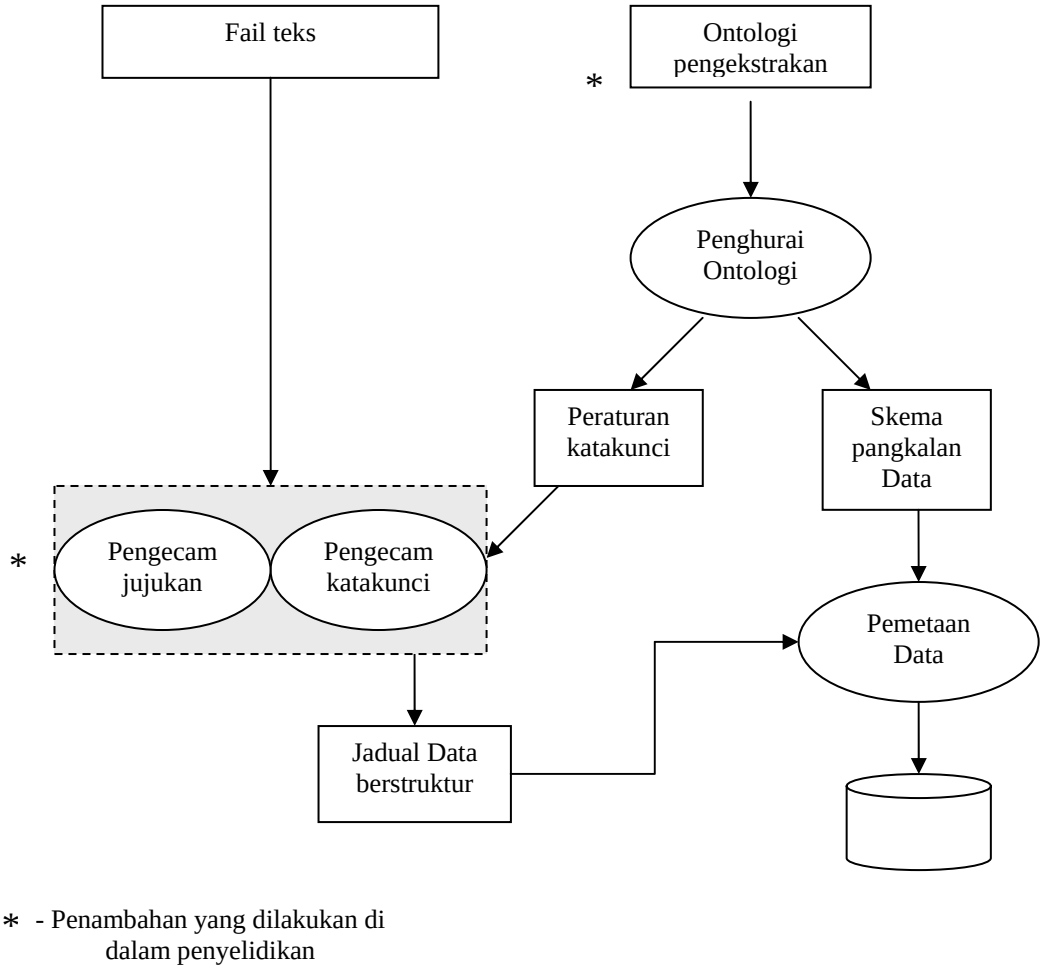
3.1 Pendahuluan

Satu reka bentuk prototaip IE berasaskan ontologi bagi domain kajian kes data hidrologi dikenali sebagai EkstrakPro akan dibincangkan di dalam bab ini. Asas reka bentuk EkstrakPro diambil daripada Embley et al. (1998). EkstrakPro terdiri daripada tiga proses iaitu proses penghuraian ontologi, proses pengecam jujukan dan kata kunci serta proses pemetaan data sebagaimana yang ditunjukkan dalam Rajah 3.1 mukasurat sebelah. Proses pengecam jujukan adalah penambahan yang dihasilkan daripada penyelidikan ini. EkstrakPro menerima dua input iaitu ontologi pengekstrakan dan data hidrologi.

Proses penghuraian ontologi akan membaca input ontologi pengekstrakan bagi menghasilkan set peraturan kata kunci dan skema pangkalan data. Manakala proses pengecam jujukan dan kata kunci akan menerima input data hidrologi. Set peraturan kata kunci yang dikehendaki akan di ekstrak dari data hidrologi berdasarkan set peraturan kata kunci berkenaan dan seterusnya maklumat-maklumat berkenaan disusun semula ke dalam jadual data berstruktur.

Berikut, proses pemetaan jadual data berstruktur ke dalam pangkalan data. Proses pemetaan data ini menghasilkan pernyataan SQL berdasarkan skema pangkalan data yang diperoleh dari proses penghuraian ontologi dan jadual data berstruktur agar dapat difahami oleh Sistem Pengurusan Pangkalan Data (DBMS). DBMS akan menyimpan data-data ke dalam medan-medan yang telah ditentukan.

Penerangan lanjut mengenai bagaimana membina ontologi pengekstrakan berserta proses-proses di dalam EkstrakPro akan dibincangkan dengan lebih terperinci di dalam bab ini.



Rajah 3.1: Reka Bentuk Embley et al.(1998) Dengan Penambahan Proses Pengecam Jujukan

3.2 Ontologi pengekstrakan

Untuk membina ontologi pengekstrakan, kajian ke atas data hidrologi berserta maklumat yang ingin di ekstrak dari data berkenaan perlu dikaji dengan teliti. Di dalam penyelidikan ini, data hidrologi yang digunakan terdapat dalam tiga format iaitu SRM, MIT dan CSV, sebagaimana yang telah dibincangkan di dalam bahagian 2.5. Manakala maklumat yang ingin di ekstrak daripada data-data hidrologi ini adalah id stesen serta nama di mana data dicerap, jenis cerapan yang dibuat, tarikh serta masa cerapan dan nilai bacaan cerapan. Seterusnya, langkah-langkah pembinaan ontologi pengekstrakan dilakukan secara manual. Berikut adalah langkah-langkah dalam menghasilkan ontologi pengekstrakan yang diringkas dari metodologi yang di cadangkan oleh Ushold dan Gruininger (1996) :

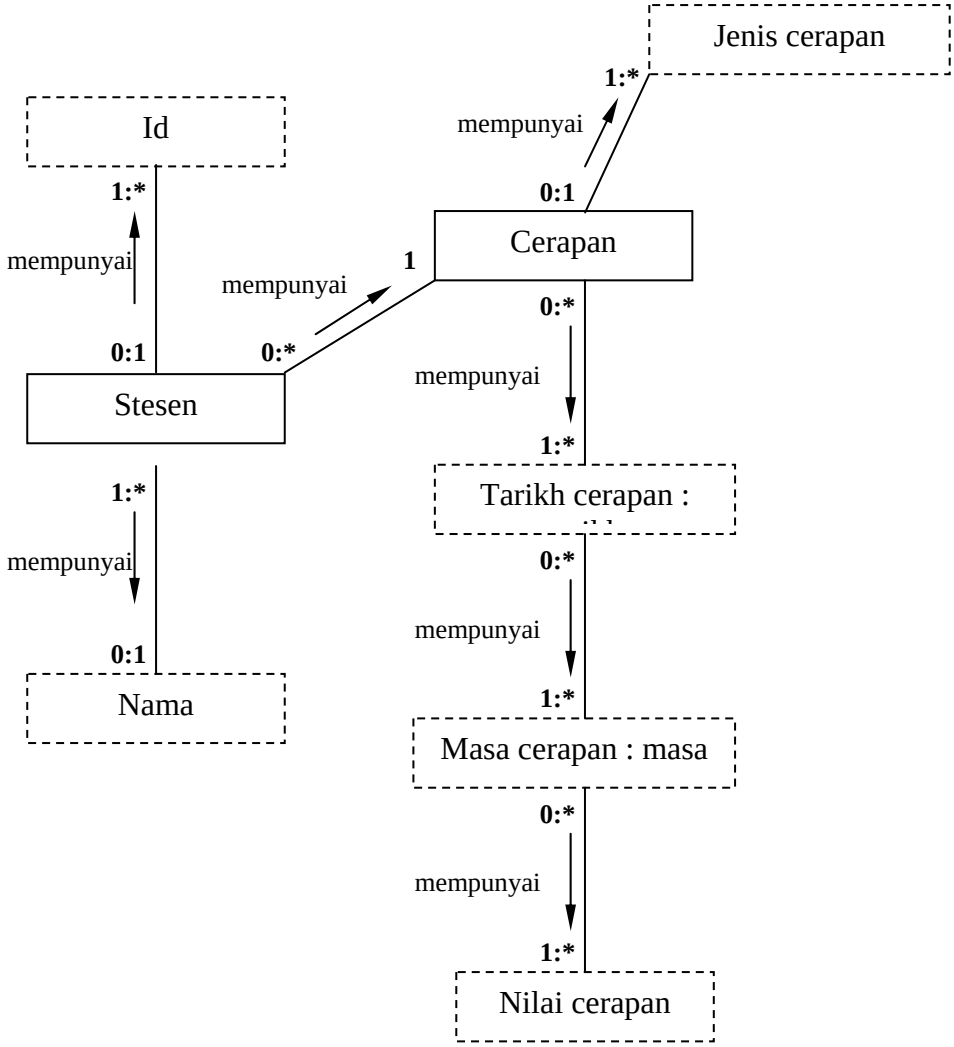
1. Mendapatkan maklumat yang ingin di ekstrak daripada objek dan hubungan di antaranya dengan menggunakan OSM
2. Menghasilkan unit objek bagi mengenal pasti corak pada maklumat yang ingin di ekstrak.

3.2.1 Penggunaan OSM

Model sistem berorientasikan objek (O*bject-oriented System Model* –OSM) digunakan untuk memberi ontologi pengekstrakan bagi domain data hidrologi. OSM diperkenalkan oleh Embley et al. (1992), mempunyai dua perwakilan grafik dan teks yang saling berkait. Ini membolehkan kita mewakilkan ontologi pengekstrakan bagi data hidrologi dalam bentuk grafik (Rajah 3.2) dan menghuraikannya ke dalam bentuk teks (Rajah 3.3).

Di dalam OSM, segi empat mewakili satu set objek. Segi empat dengan garis putus-putus mewakili set bagi objek bersifat leksikal seperti stesen_Id dan masa cerapan yang mana objek adalah perkataan yang mewakili dirinya sendiri. Manakala segi empat tanpa garis putus pula mewakili set bagi objek tidak bersifat leksikal

seperti stesen dan cerapan yang mana objek adalah pengenalpastian objek yang mewakili entiti dunia sebenar. Garis yang menghubungkan segi empat mewakili satu set hubungan. Di dalam OSM, kolon (:) selepas nama objek seperti *tarikh cerapan : tarikh*, menunjukkan bahawa objek set berkenaan adalah spesialisasi.



Rajah 3.2 : Ontologi data hidrologi JPS secara grafik

Oleh kerana bahasa persamaan untuk model telah didefinisikan untuk OSM oleh Liddle et al. (1995), dengan mudahnya model OSM secara grafik ditukarkan kepada bentuk ayat sebagaimana yang ditunjukkan dalam Rajah 3.3.

```
Stesen [0:1] mempunyai Stesen_Id [1:*];
Stesen [0:1] mempunyai Nama [1:*];
Stesen [0:1] mempunyai Cerapan [1];
…
Cerapan [0:1] mempunyai Jenis_cerapan [1:*];
Cerapan [0:*] mempunyai Tarikh_cerapan [1:*]
…
Tarikh cerapan [0:*] mempunyai Masa_cerapan[1:*];
…
Masa_cerapan [0:*] mempunyai Nilai_cerapan [1:*];
```

Rajah 3.3 : Ontologi data hidrologi JPS secara teks

3.2.2 Unit Objek

Setelah mengenal pasti objek-objek dan hubungan di antara objek bagi domain hidrologi, langkah seterusnya adalah mengenal pasti corak ke atas maklumat yang ingin di ekstrak. Beberapa contoh data hidrologi dikaji dan corak bagi setiap maklumat dikenal pasti. Untuk memudahkan proses penghuraian ontologi, unit objek (UO) diperkenalkan untuk mewakili setiap corak. Rajah 3.4 di bawah menunjukkan sintek bagi rangka UO. Ciri-ciri penting dalam corak akan diguna menghasilkan deskripsi dalam sintek berkenaan. Setiap rangka UO mempunyai nombor UO, bilangan leksikal berserta satu set sub-rangka. Nombor UO akan mewakili bilangan corak dalam satu jenis maklumat dan bilangan leksikal akan mewakili bilangan sub-rangka. Sub-rangka akan digunakan untuk memberi deskripsi kepada corak. Setiap sub-rangka boleh diwakilkan antara 2 hingga 9 deskripsi.

Nombor Unit Objek : integer

A. Bilangan item leksikal : integer

B. Nombor sub-rangka : integer
1. Nilai : Nilai yang ditetapkan atau nilai default
2. Stail : {tag, char, frasa, ayat, digit, nombor, string}
3. *Instances* : senarai string
4. Pengecualian : senarai string
5. Nombor corak : integer
   Fungsi corak : fungsi
6. Panjang Mak : integer
7. Panjang Min : integer
8. Mak : integer
9. Min : integer

Rajah 3.4 : Sintek Rangka UO

Deskripsi kelima di dalam sub-rangka bagi rangka OU iaitu nombor corak akan mempunyai fungsi corak. Terdapat lima fungsi corak yang dikenal pasti iaitu:

- **Sebarang_string** merujuk kepada apa-apa sahaja (termasuk abjad, nombor dan simbol)
- **Sebarang_digit** merujuk kepada sebarang digit
- **Sebarang_delimiter** merujuk kepada sebarang karakter khusus seperti “space bar” ataupun “tab”
- **Sebarang_tag** merujuk kepada apa yang berada di antara “< ” dan “ > ” seperti < ; >
- **Sebarang_char** merujuk kepada sebarang karakter

Langkah seterusnya adalah mengenal pasti rangka UO bagi maklumat dalam data hidrologi. Penulis telah mengambil beberapa contoh daripada tujuh jenis data hidrologi JPS untuk mengenal pasti corak bagi setiap objek yang digunakan. Terdapat sebanyak 6 jenis objek iaitu Stesen_Id, Nama_stesen, Jenis_cerapan, Tarikh_cerapan, Masa_cerapan dan Nilai_cerapan.

3.2.2.1 Stesen_Id

Daripada contoh-contoh data hidrologi yang digunakan, dapat disimpulkan bahawa stesen-Id diwakilkan oleh 7 digit sahaja. Di antara corak yang dikenal pasti adalah sebelum 7 digit bermula, adanya perkataan ‘SA-R’ ataupun ‘site’ seperti yang ditunjukkan di dalam Rajah 3.5. Rangka UO untuk Stesen_Id dijana sebagaimana yang ditunjukkan dalam Lampiran A.

1632301	1334108	2324032
site 1732001	site 1732501	site 4815001
SA-R4815001	SA-R6915111	SA-R4815001

Rajah 3.5: Contoh Stesen_Id daripada data hidrologi JPS

3.2.2.2 Nama_stesen

Berdasarkan contoh data hidrologi yang telah dipilih, data yang mempunyai corak stesen_Id ‘site 1234567’ sahaja mempunyai nama_stesen. Oleh itu untuk data hidrologi yang tidak mempunyai nama stesen, nama_stesen akan dirujuk dalam pangkalan data berdasarkan maklumat stesen_Id yang telah diperoleh.

3.2.2.3 Jenis_cerapan

Jenis cerapan juga mengalami kes yang sama iaitu tiada jenis cerapan dinyatakan di dalam data hidrologi. Maka, stesen_Id turut memainkan peranan dalam memberikan jenis cerapan dengan merujuk pangkalan data sedia ada.

3.2.2.4 Tarikh_cerapan

Corak untuk tarikh adalah berbeza-beza sebagai contoh untuk 14 hari bulan Jun tahun 2004, mungkin boleh ditulis seperti ‘14/06/2004’ atau ‘04/06/14’ atau ‘14.6.2004’ dan sebagainya. Merujuk kepada contoh data hidrologi sekali lagi beberapa corak tarikh dikenal pasti dan dinyatakan di dalam rajah 3.6.

```
20000104                20020227
SRD-10/01/2002          ERD-27/03/2002
31-DEC-2003  1-JAN-2001
```

Rajah 3.6 : Contoh Tarikh_cerapan daripada data hidrologi JPS

Untuk menjadikan sistem lebih tegar, maka setiap corak tarikh sama ada untuk data hidrologi ataupun bukan telah dikenal pasti. Rangka UO untuk tarikh cerapan telah dibina sebagaimana di dalam Lampiran B.

3.2.2.5 Masa_cerapan

Lazimnya, corak untuk masa akan melibatkan jam, minit dan saat. Ianya juga boleh ditulis dalam format 12 jam atau 24 jam. Berdasarkan contoh data, beberapa corak untuk masa cerapan telah dikenal pasti seperti yang ditunjukkan di dalam rajah 3.7. adalah unit objek yang dihasilkan berdasarkan contoh masa cerapan yang digunakan di dalam data hidrologi. Lampiran C boleh dirujuk untuk mendapatkan rangka UO bagi masa cerapan.

```
09:25               22:45               00:11
8:00:00am           3:45:10pm
SRT-14:30:15 ERT-16:21:52
122504              151722
```

R

Rajah 3.7: Contoh Masa_cerapan daripada data hidrologi JPS

3.2.2.6 Nilai_cerapan

Nilai cerapan berbeza mengikut jenis cerapan yang dilakukan. Sebagai contoh data hidrologi bagi hujan mempunyai nilai ratus bersama dua titik perpuluhan, manakala nilai untuk data hidrologi bagi aras air mempunyai nilai angka tanpa titik perpuluhan. Variasi nilai cerapan ini menyukarkan menentukan corak nilai_cerapan secara tepat. Oleh itu, pengekstrakan untuk nilai cerapan tidak dapat diwakili melalui spesifikasi objek. Nilai_cerapan boleh dikenali melalui nilai *integer* yang berturutan dalam julat semasa data jujukan.

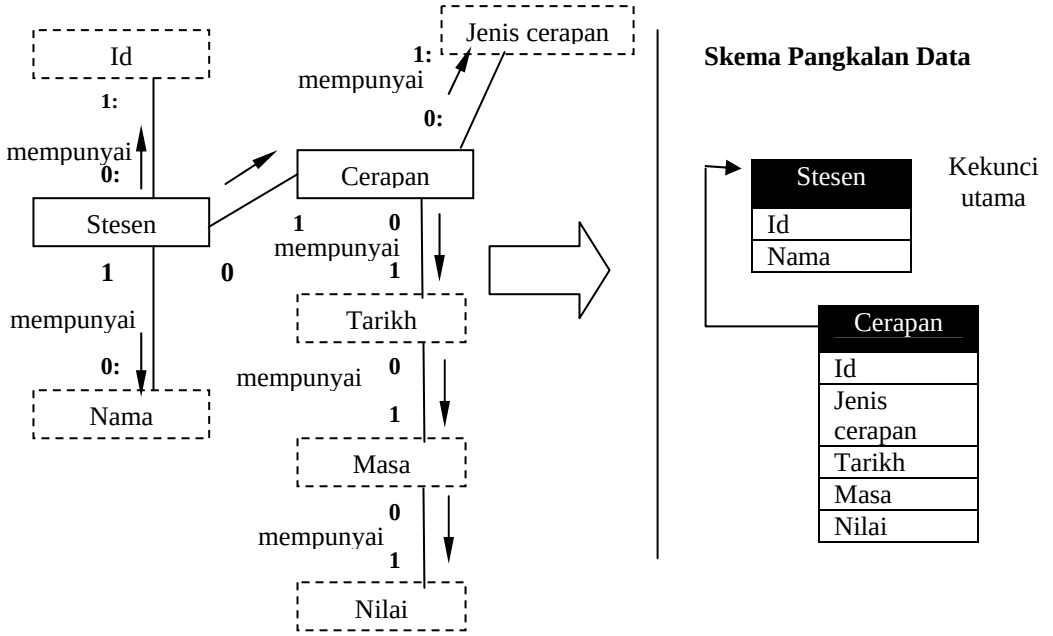
3.3 Proses Penghuraian Ontologi

Di dalam proses ini, ontologi pengekstrakan akan di hurai bagi menghasilkan set peraturan kata kunci dan skema pangkalan data. Rangka unit objek yang dihasilkan semasa ontologi pengekstrakan dibina akan menyumbang kepada set peraturan kata kunci. Setiap rangka unit objek akan menghasilkan satu kata kunci. Jika satu objek, sebagai contoh stesen_Id, mempunyai tiga rangka unit objek, maka tiga kata kunci akan dihasilkan ke dalam satu peraturan stesen_Id.

Skema pangkalan data adalah satu pernyataan SQL yang hasilkan daripada senarai nama set objek, hubungan antara objek dan kekangan. Maklumat objek-objek dan hubungannya digunakan dalam merangka struktur pangkalan data. Objek yang bersifat bukan leksikal akan mewakili jadual di dalam pangkalan data dan nama jadual akan diberi berdasarkan nama set objek tersebut. Manakala objek bersifat leksikal pula mewakili medan di dalam jadual yang mempunyai hubungan. Penormalan jadual dapat dihasilkan dengan menggunakan hubungan di antara objek bukan leksikal.

Sebagai contoh di dalam Rajah 3.8, objek *root* iaitu “stesen” mewakili satu jadual utama (*primary table*) . Objek leksikal yang mempunyai hubungan dengannya iaitu “id”dan “nama” dijadikan sebagai medan bagi jadual tersebut. Hubungan di

antara objek “stesen” dan “cerapan” menentukan “cerapan” adalah jadual kedua (*secondary table*). Seterusnya objek bukan leksikal seperti “jenis”, “tarikh”, “masa”,”nilai” menjadi medan bagi jadual “cerapan”. Oleh kerana jadual cerapan adalah jadual kedua, ia harus mempunyai satu medan yang akan menyimpan hubungan dengan kekunci di dalam jadual utama. Oleh itu “id” daripada jadual stesen akan menjadi medan di dalam jadual cerapan. Penghuraian ontologi pengekstrakan dalam menghasilkan skema pangkalan data diringkaskan di dalam Rajah 3.8.



Rajah 3.8 : Skema pangkalan data daripada ontologi pengekstrakan

Proses penghuraian ontologi akan menggunakan ontologi pengekstrakan untuk menghasilkan set peraturan kata kunci dan skema pangkalan data. Set peraturan kata kunci diperoleh daripada nama U, manakala peraturan pemadanan kata kunci akan merujuk kepada sub-rangka di dalam unit objek.

3.4 Proses Pengecam Jujukan

Pendekatan ontologi sebagaimana yang dicadangkan oleh BYU, akan membuat perbandingan antara data-data di dalam fail dengan kata kunci. Sekiranya data menepati kata kunci yang diperoleh, maka data tersebut akan dimasukkan ke dalam jadual yang telah ditetapkan di dalam pangkalan data. Rajah 3.9 menunjukkan algoritma pengekstrakan data bagi EkstrakPro sebagaimana yang dicadangkan oleh BYU.

```
1.  Baca fail input
2.  WHILE not EOF DO
    {
3.      Baca Baris & Dapatkan current.data
4.      For Bil_KataKunci = 1 to MaxKataKunci
5.      {
6.          Bandingkan KataKunci
7.          IF current.data = KataKunci

8.          {
9.              Masukkan nilai current.data ke DB

10.             Bil_KataKunci = MaxKataKunci

11.         }
12.         ELSE
13.             Bil_KataKunci = Bil_KataKunci +1

14.     }
15.  }
```

Rajah 3.9 : Algoritma EkstrakPro

Baris pertama adalah proses membaca fail input yang ingin di ekstrak. Baris kedua adalah proses pengulangan sehingga akhir fail input terbabit. Baris keempat pula merupakan bacaan ke atas data secara baris ke baris. Baris kelima dalam algoritma tersebut iaitu pengecaman data menggunakan kata kunci yang mana kata kunci dihasilkan daripada penghuraian ontologi. Manakala baris keenam dan ketujuh akan memasukkan data yang telah dikenal pasti ke dalam jadual pangkalan data yang telah ditetapkan semasa proses penguraian ontologi.

Satu masalah yang timbul daripada penggunaan algoritma di atas adalah dari segi masa. Jangka masa proses pengecaman bergantung kepada saiz sesebuah fail bersama dengan bilangan kata kunci.

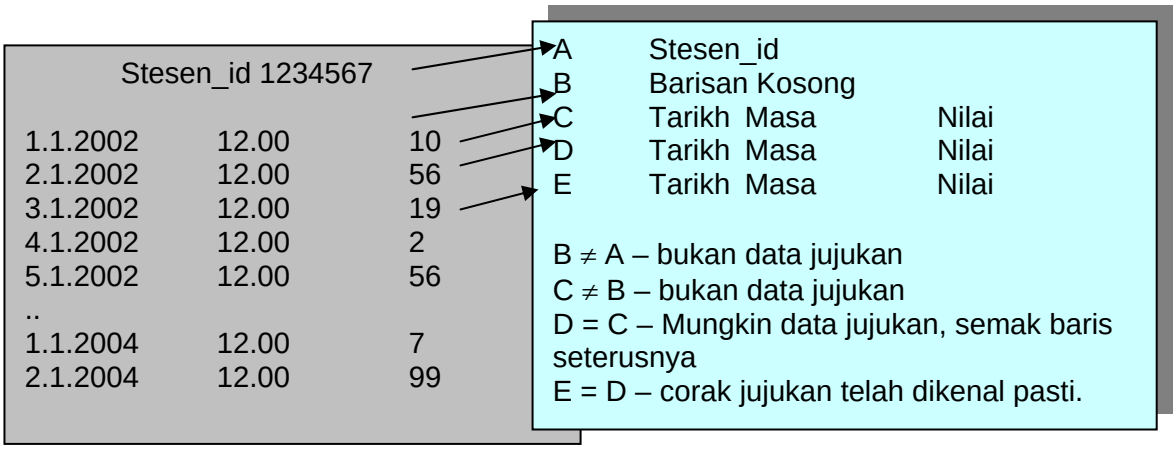
Tujuan proses pengecaman jujukan adalah untuk mengelakkan proses pengecaman kata kunci yang berulang -ulang. Merujuk kepada data hidrologi dalam contoh di Rajah 3.10, didapati lajur pertama menyimpan maklumat bagi nilai tarikh, lajur kedua menyimpan nilai masa dan lajur ketiga menyimpan nilai bacaan. Dengan memperkenalkan algoritma pengecaman jujukan, proses kata kunci tidak perlu dilakukan ke atas setiap baris data input. Algoritma ini berfungsi untuk mengenal pasti corak susunan jujukan dalam data berkenaan. Setelah corak jujukan dikenal pasti, proses memasukkan data ke pangkalan data akan dijalankan secara automatik tanpa perlu melakukan pengecaman kata kunci pada baris berikutnya.

```
Stesen_id 1234567

1.1.2002        12.00        10
2.1.2002        12.00        56
3.1.2002        12.00        19
4.1.2002        12.00        2
5.1.2002        12.00        56
..
1.1.2004        12.00        7
2.1.2004        12.00        99
```

Rajah 3.10 : Corak jujukan data hidrologi JPS

Secara ringkas, algoritma pengecaman jujukan bertindak dengan cara membandingkan nilai maklumat bagi setiap lajur di antara baris-baris. Sekiranya nilai maklumat untuk baris-baris (sekurang-kurangnya 3 baris) yang dibandingkan adalah sepadan maka, untuk baris-baris berikutnya, nilai lajur telah ditentukan oleh corak jujukan yang dikenal pasti (Rajah 3.11).



Rajah 3.11: Notasi algoritma pengecaman jujukan

Algoritma pengecaman jujukan adalah seperti Rajah 3.12 di bawah. Algoritma EkstrakPro diperbaiki dengan memasukkan algoritma pengecaman jujukan ke dalamnya sebagaimana yang ditunjukkan dalam Rajah 3.13.

```
Baca fail Input
 While not EOF
          {
                    Baca baris ;
                    Kenalpasti dengan peraturan kata kunci;
                    Simpan nilai kata kunci dalam fail new.keyword;
                    Buat perbandingan dengan previous.keyword;
                    If true {
                    i = i +1;
                    }
                    If i > 3 then {
                              Simpan nilai new.keyword ke dalam
                              PeraturanJujukan.keyword;
                    }
                    else if
                    {
                              new.keyword diumpukan kepada
                              previous.keyword;
                    }
}
```

Rajah 3.12 : Algoritma pengecam jujukan

```
1.  Baca fail input

2.  WHILE not EOF DO

    {
3.  Baca Baris & Dapatkan current.data
4.  Panggil Algoritma jujukan
5.  IF jujukan dikenalpasti == TRUE

6.  {
7.      Masukkan current.data ke DB

8.  }
9.  ELSE

10. {
11.     For Bil_KataKunci = 1 to MaxKataKunci

12.     {
13.      Bandingkan KataKunci

14.      IF current.data = KataKunci

15.      {
16.          Masukkan nilai current.data ke DB

17.          Bil_KataKunci = MaxKataKunci

18.      }
19.      ELSE

20.          Bil_KataKunci = Bil_KataKunci +1

21.     }
22.  }
23.  }
```

Penambahan Algoritma pengecaman jujukan di dalam algoritma pengekstrakan data

Rajah 3.13 : Algoritma EkstrakPro dengan Algoritma jujukan

3.5 Proses Pemetaan

Di dalam proses pemetaan, kata kunci yang di ekstrak akan dipadankan dengan skema SQL untuk memplotkan rekod di dalam skema pangkalan data. Proses ini menghubungkan jadual data berstruktur dengan skema pangkalan data sebelum disimpan ke dalam pangkalan data. Sebagaimana yang telah ditunjukkan dalam Rajah 3.8 di atas, skema pangkalan data iaitu skema SQL mengandungi dua jadual

iaitu “stesen” dan “cerapan”. Pemetaan ini akan menghasilkan skrip *insert statement*, *standard database query language* (SQL).

3.6 Pengujian

Pengujian dilaksanakan untuk menguji ketahanan pengekstrakan dan kebolehan algoritma pengecam jujukan mengurangkan masa pengekstrakan. Pengekstrakan diuji dengan data hidrologi JPS iaitu data taburan hujan, penyejatan, ketinggian air sungai, enapan terapung dan kualiti air. Bagi menguji ketahanan pengekstrakan, sampel data ujian di ekstrak menggunakan Sistem EkstrakPro. Ketepatan data diplotkan ke dalam pangkalan data menjadi ukuran ketahanan di dalam pengujian ini. Selain menggunakan sampel data ujian, Sistem EkstrakPro juga diuji dengan data yang diubah struktur data.

Pengujian kedua yang dilaksanakan adalah untuk melihat keupayaan algoritma pengecam jujukan rekod mengurangkan masa pengekstrakan. Ini dapat dilihat dengan membandingkan masa pemprosesan pengekstrakan yang menggunakan algoritma pengecam jujukan dengan pengekstrakan tanpa algoritma pengecam jujukan.

3.7 Kesimpulan

Secara kesimpulannya, metodologi penyelidikan merangkumi proses-proses iaitu membina ontologi pengekstrakan, penghuraian ontologi, pengecam jujukan rekod dan pengujian. Jadual 3.1 menunjukkan input, teknik, output dan sumbangan daripada proses metodologi penyelidikan. Selain daripada itu, reka bentuk asas prototaip EkstrakPro turut dibincangkan bagi membolehkan proses pengimplimentasian dilakukan dengan mudah. EkstrakPro terdiri daripada tiga proses utama iaitu proses penghuraian ontologi, proses pengecam jujukan dan kata kunci serta proses pemetaan data.

Jadual 3.1 : Ringkasan metodologi penyelidikan

	INPUT	***TEKNIK***	***OUTPUT***	***SUMBANGAN ILMIAH***
1.Membina ontologi pengekstrakan	Data hidrologi JPS	OSM, (Embley et al.,1992)	Ontologi pengekstrakan data hidrologi	Ontologi pengekstrakan bidang data hidrologi JPS + Unit Objek
2.Membina penghuraian ontologi	Ontologi pengekstrakan	Algoritma (Embley et al., 1998)	Set peraturan kata kunci + Skema pangkalan data	
3.Membina algoritma pengecam jujukan rekod	Set peraturan kata kunci	Cadangan Penyelidikan	Jadual data berstruktur	Algoritma Pengecaman Jujukan rekod
4.Pengujian	Data hidrologi JPS	Empirikal	- Hasil data yang diplotkan ke dalam pangkalan data -Perbandingan masa pemprosesan	

BAB 4

IMPLIMENTASI

4.1 Pendahuluan

Bagi menguji keberkesanan pengekstrakan maklumat berasaskan ontologi bagi domain hidrologi, satu prototaip pengekstrakan data iaitu EkstrakPro telah dibangunkan. Tujuan utama prototaip EkstrakPro dibina adalah untuk melakukan proses pengujian dan sekali gus membukti ketahanan sistem pengekstrakan data berasakan ontologi dalam domain hidrologi. Antara muka pengguna dibina bagi memudahkan pengguna dalam memanipulasi sistem prototaip berkenaan.

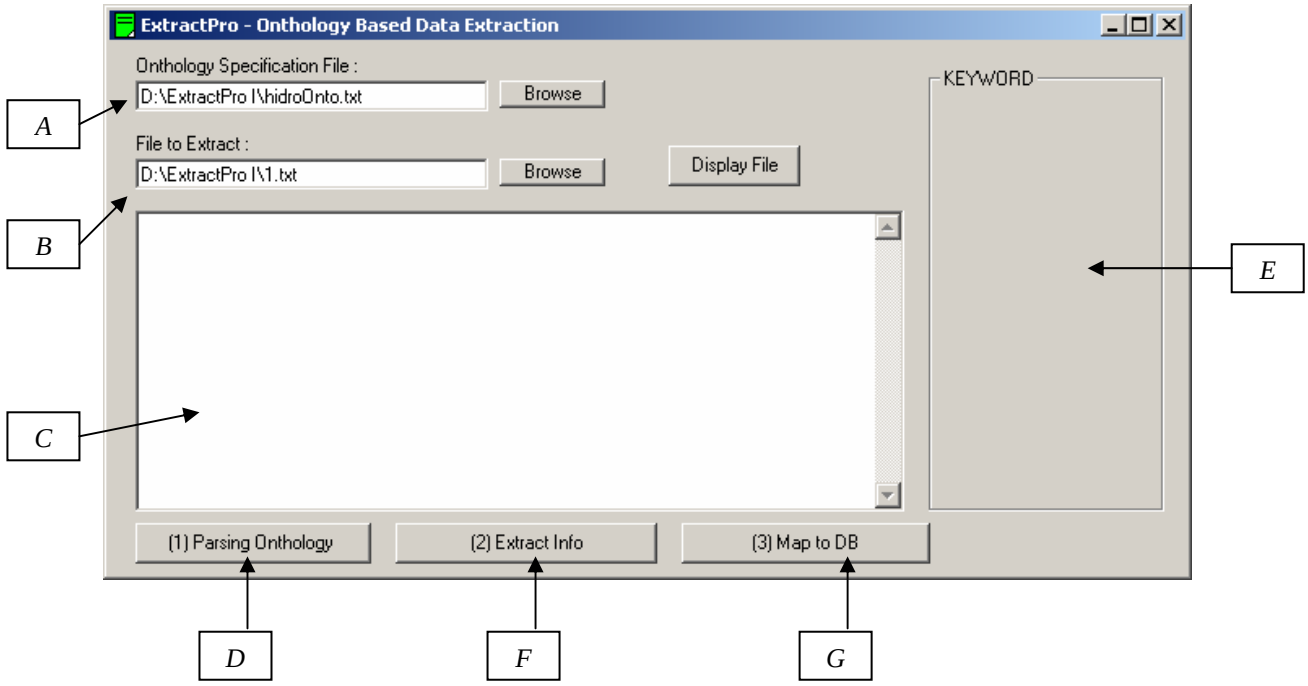
4.2 Spesifikasi Sistem

Pembangunan prototaip EkstrakPro dilakukan dengan spesifikasi berikut :

Sistem OS : Microsoft Window XP
Bahasa Pengaturcaraan : Visual Basic & SQL
Pangkalan Data : Microsoft Access

4.3 Antara Muka Sistem

Satu antara muka pengguna telah dibangunkan bagi memudahkan pengguna memasukkan kedua-dua input iaitu ontologi pengekstrakan dan data hidrologi. Selain itu, antara muka membenarkan maklumat data dan senarai kata kunci dipaparkan. Pengguna juga boleh melaksanakan proses pengekstrakan dengan mengendalikan butang-butang yang telah disediakan. Rajah 4.1 berikut menunjukkan antara muka EkstrakPro bersama fungsi butang-butang di dalamnya. Manakala, saling kaitan di antara reka bentuk prototaip EkstrakPro dengan antara muka yang telah direka dapat dilihat dengan jelas di dalam Rajah 4.2.



Rajah 4.1 : Antara muka EkstrakPro

Ringkasan penerangan bagi setiap butang yang di label pada Rajah 4.1 adalah seperti berikut:

A ***Kontrol Dialog*** - Input lokasi dan nama ontologi pengekstrakan.

B ***Kontrol Dialog*** - Input lokasi dan nama data teks.

C ***Paparan maklumat data*** – Memaparkan kandungan fail yang dibaca.

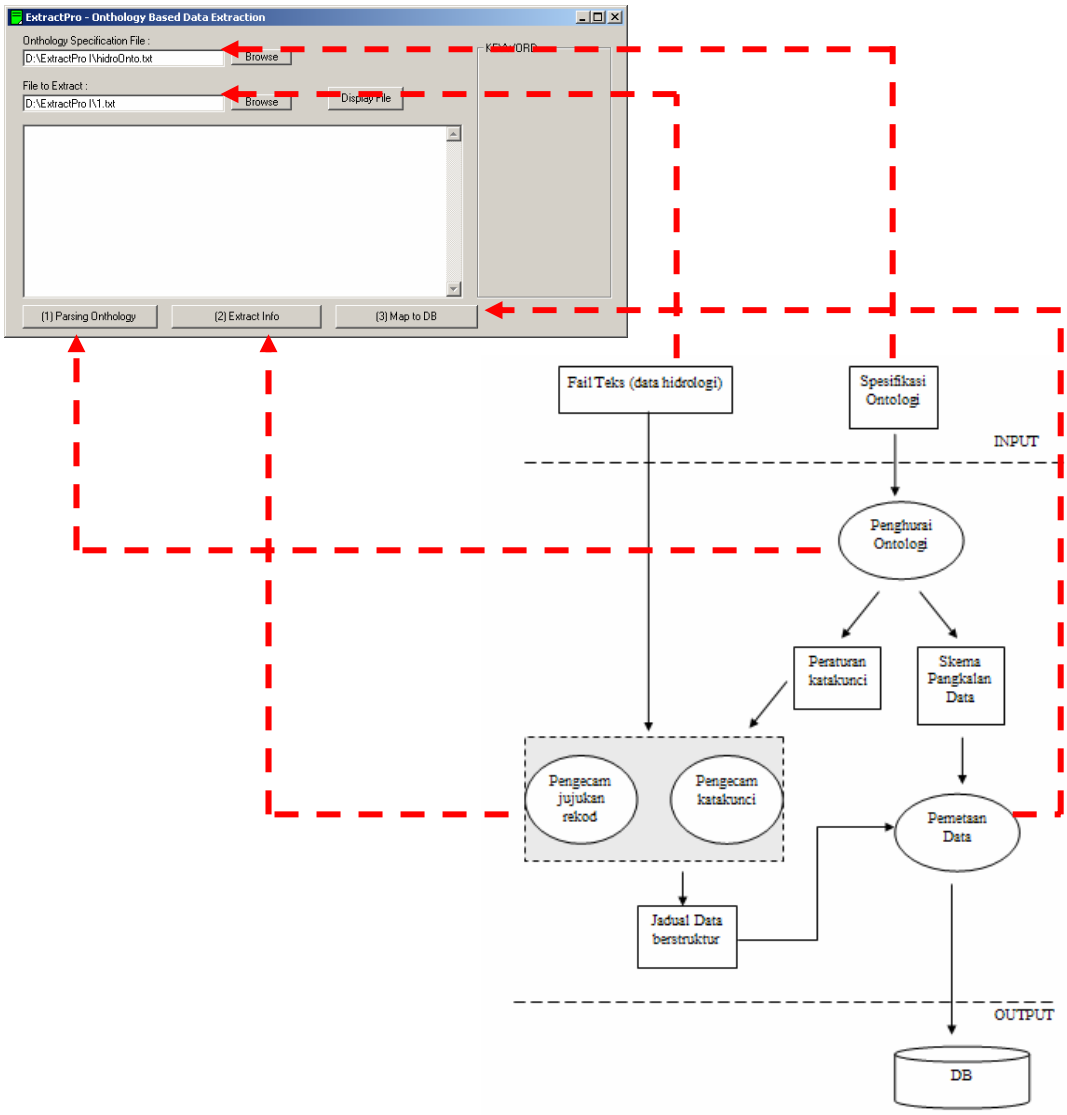
D ***Penghuraian Ontologi*** – Proses menghasilkan kata kunci dan hubungan kata kunci daripada ontologi pengekstrakan.

E ***Paparan kata kunci*** – Senarai kata kunci dipaparkan di dalam ruangan ini apabila penghuraian ontologi di laksanakan.

F ***Mengekstrak maklumat*** – Proses memadankan kata kunci dengan data dan perlaksanaan pengecaman jujukan rekod.

G ***Butang pemetaan data*** - Memetakan data yang telah di ekstrak ke dalam pangkalan data.

(a) Antara muka EkstrakPro



(b) Reka bentuk Algoritma Pengekstrakan Data

Rajah 4.2 : Reka Bentuk Sistem dan Antara Muka Prototaip EkstrakPro

4.3 Implementasi Proses Penghuraian Ontologi

Fungsi penghuraian ontologi adalah untuk menghasilkan kata kunci dan skema pangkalan data daripada maklumat fail input ontologi pengekstrakan. Fail input bagi ontologi pengekstrakan dihasilkan daripada objek dan hubungan di antaranya bersama-sama senarai UO yang dikenal pasti. Rajah 4.3 memberikan satu contoh fail input ontologi pengekstrakan bagi maklumat Tarikh cerapan.

```
Maklumat Tarikh cerapan
Bilangan corak 3
####[1960 to 2111]##[01 to 12]##[01 to 31]
“SRD” | “ERD””-“##[01 to 31]”/”##[01 to 12]”/”####[1960 to 2111]
##[01 to 31]”-“”JAN” | “FEB” | “MAC” | “APR” | “MAY” | “JUN” | “JULY” | “AUG” | “SEP” | “OCT” | “NOV” |
“DEC””-”####[1960 to 2111]
```

Rajah 4.3 : Input Ontologi pengekstrakan bagi Tarikh Cerapan

Atur cara dimulakan dengan membaca fail ontologi pengekstrakan yang bagi setiap maklumat yang ingin di ekstrak. Bilangan corak bergantung pada bilangan UO yang telah dihasilkan. Setiap baris (bermula dari baris ketiga) dalam fail input ontologi pengekstrakan akan mewakili satu jenis corak. Sekiranya terdapat corak yang baru, pengguna perlu membina UO bagi corak berkenan dan memasukkannya ke dalam fail input sebagai baris yang baru. Rajah 4.4 di sebelah menunjukkan keratan atur cara yang membaca fail input ontologi pengekstrakan dan menyimpannya sebagai satu set kata kunci.

```
Open txtFileName.Text For Input As FP1
Set DB1 = OpenDatabase(DBName, False, False)
Set RS1 = DB1.OpenRecordset("dt_real") 'this opens the whole
table
d = 1

FP1 = FreeFile
'On Local Error GoTo ER1
Open txtFileName.Text For Input As FP1

Do Until EOF(FP1) '
Line Input #FP1, s
'RS1.AddNew

MyPos = InStr(s, "site")
If MyPos > 0 Then
   s1 = Mid$(s, MyPos + 5, 6)
'    txtResults.Text = txtResults.Text & s1 & CRLF
'    RS1.AddNew
'    RS1.Fields(0).Value = s1
'    RS1.Update
   End If

MyPos1 = InStr(s, "Year")
If MyPos > 0 Then
   s2 = Mid$(s, MyPos1 + 5, 4)
   RS1.AddNew

  ' txtResults.Text = txtResults.Text & s2 & CRLF
   'txtResults.Text = txtResults.Text & sArray(i) & CRLF
      ' RS1.AddNew
      ' RS1.Fields(1).Value = sArray(i)
       'RS1.Update
   End If
sArray = Split(s)
m = 0
Dim nm
For i = LBound(sArray) To UBound(sArray)
   MyCheck = sArray(i) Like "*[.]##"

   If MyCheck = True Then
      RS1.AddNew
      RS1.Fields(0).Value = s1
      m = m + 1
      If Not m = 13 Then
      RS1.Fields(2).Value = d & "/" & m & "/" & s2
      RS1.Fields(6).Value = sArray(i)
      RS1.Update
      End If
      If m = 12 Then
      d = d + 1
      End If
```

Rajah 4.4 : Keratan Atur Cara Penghuraian ontologi bagi menghasilkan peraturan kata kunci

Selain set peraturan kata kunci, penghuraian ontologi juga menghasilkan skema pangkalan data dengan menggunakan objek utama sebagai nama jadual dan objek inheren menjadi medan di dalam jadual. Skema struktur pangkalan data ini dibina menggunakan bahasa SQL. Rajah 4.5 di sebelah menunjukkan keratan atur cara bagi skema struktur pangkalan data yang dijana daripada ontologi pengekstrakan.

```
Create table Stesen (
        Stesen_ID integer
        Nama varchar (80),
        Jenis_bacaan (50),
)

Create table Cerapan (
        Stesen_ID integer
        Tarikh varchar(10),
        Masa varchar(10),
        Bacaan real (10)
)
```

Rajah 4.5 : Contoh Skema Pangkalan Data

4.4 Implementasi bagi Proses Pengecaman Jujukan dan Kata Kunci

Set kata kunci yang dihasilkan oleh penghuraian ontologi akan digunakan dalam proses pengekstrakan maklumat daripada data input. Pemadanan set kata kunci dilaksanakan ke atas setiap aksara di dalam data input. Aksara yang telah dikenal pasti sebagai kata kunci akan di ekstrak dan disimpan di dalam jadual data berstruktur. Jadual ini mengandungi data yang di ekstrak dan maklumat pemetaannya. Rajah 4.6 adalah keratan atur cara pengekstrakan maklumat.

```
For i = LBound(sArray) To UBound(sArray)
   MyCheck = sArray(i) Like "*[.]##"

   If MyCheck = True Then
      RS1.AddNew
      RS1.Fields(0).Value = s1
      m = m + 1
      If Not m = 13 Then
      RS1.Fields(2).Value = d & "/" & m & "/" & s2
      RS1.Fields(6).Value = sArray(i)
      RS1.Update
      End If
      If m = 12 Then
      d = d + 1
      End If
   End If
Next
```

Rajah 4.6 : Keratan Atur Cara Pengekstrakan Kata kunci

4.5 Implementasi bagi Proses Pemetaan Data

Data daripada jadual berstruktur akan dipetakan ke dalam pangkalan data menggunakan yang pemetaan data jadual berstruktur dan skema pangkalan data. Maklumat pemetaan akan dipadankan dengan nama medan di dalam skema. Pernyataan *insert* akan dihasilkan untuk dilarikan di dalam sistem pangkalan data Microsoft Access. Rajah 4.7 menunjukkan kenyataan *insert* yang telah dihasilkan.

```
Insert into dt_integer (fid,tid,did,mcode,qcode,val)
        values (“1222352”, “20030204113000”, “”, “101”,”201”,”50080”)
    Insert into dt_integer (fid,tid,did,mcode,qcode,val)
        values (“1222352”, “20030204160611”, “”, “101”,”201”,”50083”)
    Insert into dt_integer (fid,tid,did,mcode,qcode,val)
        values (“1222352”, “20030204180316”, “”, “101”,”201”,”50085”)
    Insert into dt_integer (fid,tid,did,mcode,qcode,val)
        values (“1222352”, “20030205135505”, “”, “101”,”201”,”50078”)
```

Rajah 4.7 : Keratan Pernyataan *Insert*

4.6 Kesimpulan

Implementasi rangka kerja IE berasaskan ontologi telah dilaksanakan. Setiap proses rangka kerja, telah diprogramkan dan hasilnya prototaip EkstrakPro. Dengan input data teks hidrologi JPS dan ontologi pengekstrakan, maklumat yang telah di ekstrak akan dipetakan ke dalam sistem pangkalan data *Microsoft Access*. Sistem prototaip ini membolehkan penyelidikan menjalankan pengujian keberkesanan IE berasaskan ontologi mengekstrak data teks hidrologi.

BAB 5

PENGUJIAN

5.1 Pendahuluan

Proses pengujian merupakan proses kritikal dalam membuktikan penyelesaian yang dipilih adalah bersesuaian dengan masalah penyelidikan. Di dalam bab ini, dua pengujian dijalankan ke atas sistem EkstrakPro, yang pertama adalah untuk menguji ketahanan pengekstrakan maklumat dan kedua adalah untuk menguji masa pengekstrakan. Ketahanan sesebuah IE ditentukan melalui keupayaannya mengekstrak maklumat dari pelbagai jenis data dan menyimpannya ke dalam pangkalan data dengan tepat. Masa pengekstrakan maklumat pula ditentukan bermula dari perbandingan kata kunci sehingga pemetaan maklumat ke dalam pangkalan data.

5.2 Penyediaan Data Ujian

Data ujian terbahagi kepada tiga set data hidrologi. Pertama adalah data hidrologi yang diperoleh dari JPS. Ia merangkumi data-data cerapan bagi taburan hujan, penyejatan, aras air sungai, enapan terapung sungai dan kualiti air sungai. Sebanyak 100 data hidrologi diambil secara rawak daripada 20 stesen cerapan di seluruh Malaysia. Format bagi data-data ini adalah SRM, MIT dan CSV. Beberapa

contoh keratan data hidrologi yang digunakan bagi kategori pertama boleh didapati di dalam Lampiran D.

Manakala kategori kedua adalah data hidrologi rekaan. Set data ini diperoleh dengan melakukan perubahan ke atas data JPS sedia ada bagi mendapat format data yang baru. Beberapa perubahan mudah dilakukan seperti mengubah kedudukan jujukan data, menambah dan membuang jujukan data serta menambah maklumat-maklumat tambahan dalam *header*. Selain daripada itu, data-data baru juga diperoleh daripada kutipan dari internet dan juga dihasilkan sendiri oleh penulis. Sebanyak 30 set data baru telah dihasilkan. Contoh keratan data hidrologi yang digunakan bagi kategori kedua boleh didapati di dalam Lampiran E.

Set data ujian ketiga merupakan data hidrologi yang berlainan saiz namun terdiri daripada satu format iaitu CSV sahaja disediakan. Julat baris yang terkandung dalam fail input adalah di antara 200 hingga ke 1800 baris data. Beberapa keratan contoh data hidrologi daripada set ketiga disertakan di dalam Lampiran F.

5.3 Ujian Ketahanan Pengekstrakan maklumat

Pengujian ketahanan pengekstrakan maklumat bagi sistem EkstrakPro dilakukan dalam dua peringkat. Peringkat pertama adalah perbandingan sistem EkstrakPro dengan sistem *MHIS Dataload*. Perbandingan ini bertujuan untuk membuktikan keupayaan sistem EkstrakPro mengekstrak data adalah setara dengan sistem *MHIS Dataload*. Pengujian dilakukan ke atas set data ujian pertama dan ketepatan pengekstrakan maklumat bagi kedua-dua sistem berkenaan dikenal pasti. Peratus ketepatan diukur berdasarkan rumus di bawah.

$$\text{Peratus ketepatan} = \frac{\text{Bil. maklumat tepat yang telah diekstrak}}{\text{Bil. Maklumat tepat sebenar}} \text{ X } 100$$

Jadual 5.1 di sebelah menunjukkan keputusan pengujian ketahanan pengekstrakan maklumat peringkat pertama. Daripada pengujian ini, didapati keputusan kedua-dua sistem pengekstrakan maklumat secara puratanya hampir mencapai ketepatan 100 %. Namun dengan hanya menggunakan satu algoritma, EkstrakPro berasaskan ontologi dapat mengekstrak maklumat dari pelbagai jenis data hidrologi. Berbeza dengan sistem *MHIS Dataload* yang memerlukan algoritma berbeza untuk setiap jenis data.

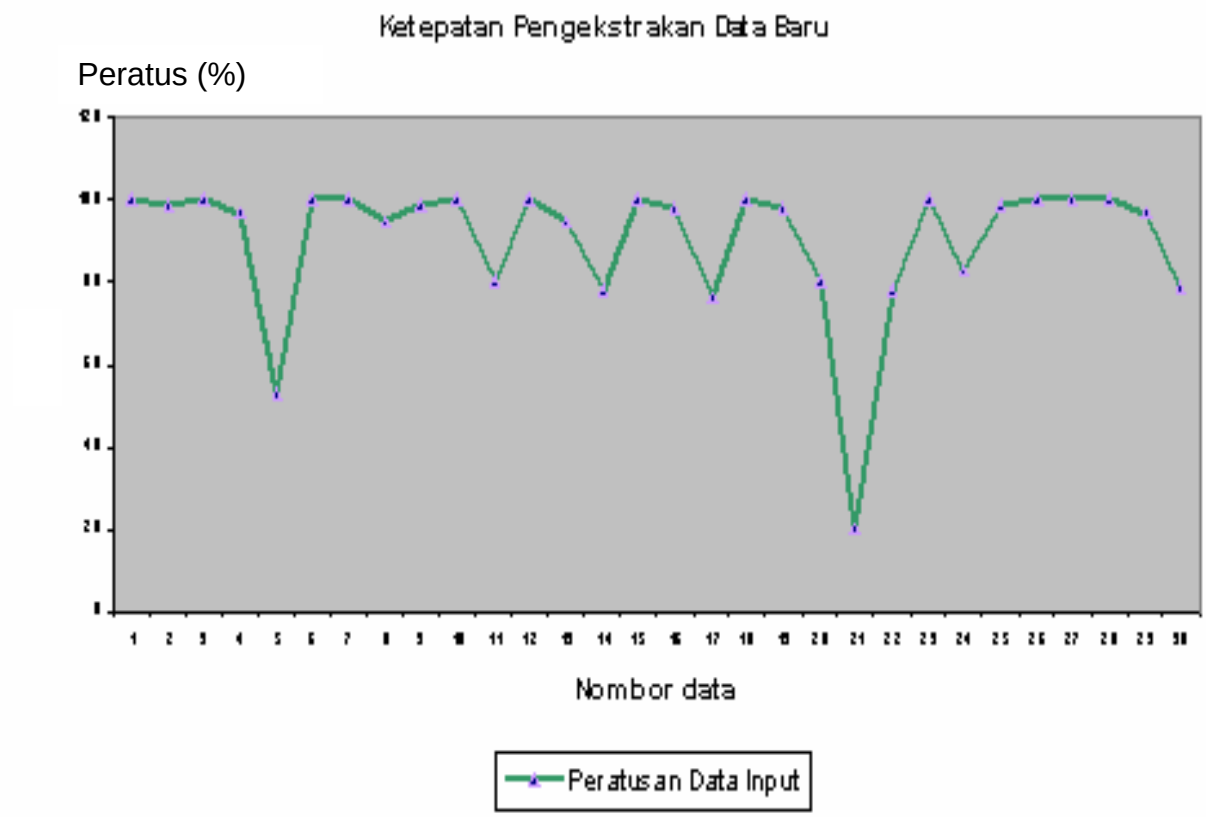
Jadual 5.1 : Peratus ketepatan bagi algoritma *MHIS Dataload* dan algoritma EkstrakPro

	Peratus ketepatan	Peratus ketepatan
Taburan hujan	Algo. Taburan Hujan MHIS ~ 100%	Algo. Ontologi ~ 100%
Penyejatan	Algo. Penyejatan ~ 100%	Algo. Ontologi ~ 100%
Ketinggian Air Sungai	Algo. Ketinggian air sungai ~ 100%	Algo. Ontologi ~ 100%
Enapan Terapung	Algo. enapan terapung ~ 100%	Algo. Ontologi ~ 100%
Kualiti Air	Algo Kualiti air ~ 100%	Algo. Ontologi ~ 100%

Satu lagi kelebihan utama menggunakan pengekstrakan maklumat berasaskan ontologi adalah ianya mampu mengekstrak format data baru tanpa perlu melibatkan proses pengaturcaraan semula. Untuk membuktikannya, pengujian ketahanan peringkat kedua dilaksanakan. Pengujian dilakukan ke atas set data ujian kedua yang melibatkan data hidrologi rekaan. Rajah 5.1 menunjukkan keputusan ujian peringkat kedua.

Rajah 5.1 menunjukkan sebanyak 70% data hidrologi rekaan dapat di ekstrak dengan tepat. Baki sebanyak 30% tidak dapat di ekstrak dengan tepat kerana kekurangan dari segi perbandingan kata kunci. Sebagai contoh, untuk data input nombor 21 mempunyai nilai tarikh yang disusun dalam kedudukan Januari, 23, 1999. Oleh kerana tiada unit objek yang menerangkan corak bagi tarikh ini, maka semasa perbandingan kata kunci dilakukan, data ini tidak akan dikenal pasti sebagai tarikh.

Ini menyebabkan data yang di ekstrak tidak tepat. Cara terbaik untuk mengurangkan ralat seperti ini berlaku adalah dengan memastikan unit objek untuk setiap objek yang ingin di ekstrak adalah lengkap.



Rajah 5.1 : Peratus ketepatan pengekstrakan maklumat terhadap jenis data

5.4 Ujian Masa Pengekstrakan maklumat

Tujuan utama pengujian ini dilakukan adalah untuk membuktikan algoritma pengecaman jujukan yang dihasilkan dapat mengurangkan masa pengekstrakan. Pengujian masa pemprosesan ini turut dilakukan dalam dua ukuran. Ukuran pertama menggunakan notasi-O untuk mengira masa algoritma pengekstrakan dalam sistem EkstrakPro. Ukuran kedua akan menggunakan masa pengekstrakan yang dilakukan ke atas set data ketiga.

Notasi-O digunakan untuk menilai sesebuah algoritma yang dibina dari segi masa dan saiz ruang pelaksanaannya. Pengujian kali ini menggunakan notasi-O untuk membandingkan antara algoritma EkstrakPro (Rajah 3.9) dengan algoritma EkstrakPro bersama algoritma jujukan (Rajah 3.13). Dengan beranggapan panjang fail data hidrologi sebagai n dan bilangan kata kunci sebagai m, masa pengekstrakan untuk algoritma EkstrakPro adalah O(mn). Manakala untuk algoritma EkstrakPro bersama algoritma jujukan pula, masa pengekstrakan bergantung pada tiga kes seperti berikut :

1. Kes ideal

Kes ideal diperoleh sekiranya data hidrologi yang di ekstrak mempunyai data-data jujukan dari permulaan fail. Ini bermakna proses perbandingan kata kunci hanya akan dilakukan pada tiga baris pertama sahaja, manakala baris-baris seterusnya akan dilakukan tanpa melakukan perbandingan. Notasi-O bagi kes ideal adalah O(n).

2. Kes sederhana

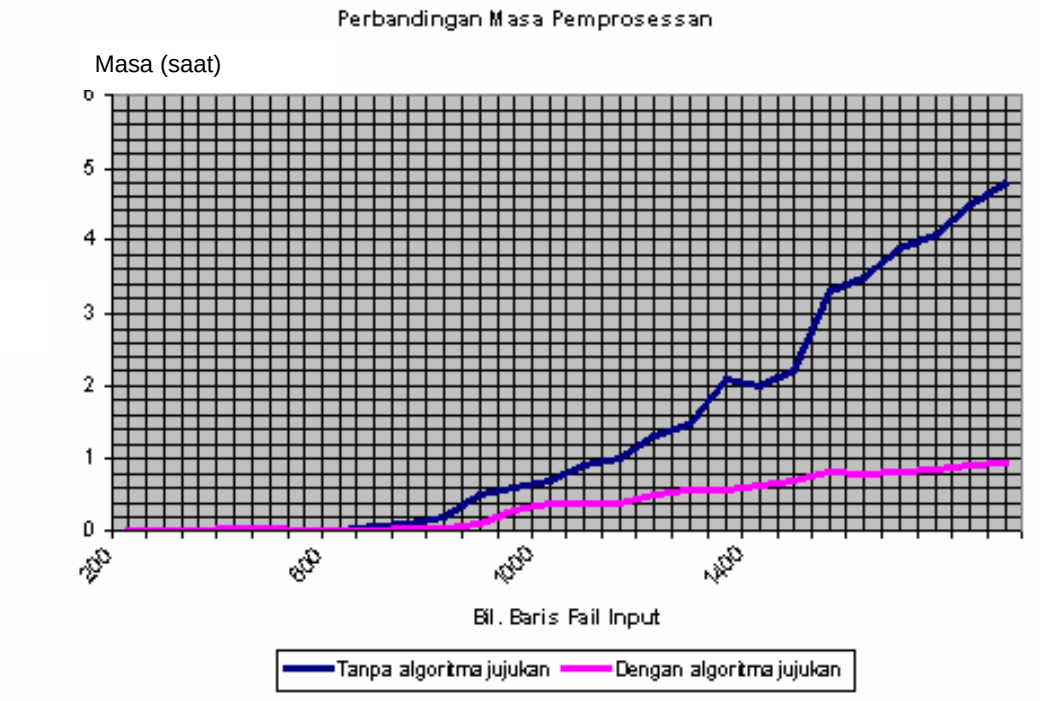
Kes sederhanaan terjadi apabila data input yang digunakan mempunyai *header* di permulaan fail. Ini menyebabkan proses perbandingan kata kunci dilaksanakan pada setiap baris dan data dalam *header* sehingga algoritma menjumpai data jujukan. Dengan beranggapan saiz *header* sebagai h, notasi-O bagi kes sederhana adalah O(n)+h. Ini bermakna semakin besar saiz *header* semakin lama proses pengekstrakan. Namun, sekiranya data jujukan adalah bersaiz besar, maka penggunaan algoritma pengecaman jujukan akan membantu mengurangkan masa pemprosesan walaupun fail input mempunyai data *header*.

3. Kes tidak ideal

Kes tidak ideal berlaku sekiranya algoritma tidak menjumpai sebarang data jujukan dalam fail input. Ini menyebabkan masa pengekstrakan meningkat kerana penambahan semakan daripada algoritma jujukan. Namun untuk penyelidikan kali ini, sebagaimana yang telah dinyatakan dalam skop bahawa

data input yang digunakan adalah data hidrologi berjujukan, maka kes tidak ideal tidak akan berlaku.

Ujian kedua menggunakan masa pengekstrakan maklumat, yang mana masa pengekstrakan dikira bermula dari perbandingan kata kunci sehingga pemetaan maklumat ke dalam pangkalan data. Set data ujian ketiga digunakan untuk tujuan berbandingan. Daripada Rajah 5.2 didapati algoritma EkstrakPro yang menggunakan penambahan algoritma pengecaman jujukan dapat mempercepatkan lagi masa pengekstrakan. Untuk data yang bersaiz kecil iaitu kurang daripada 600 baris, didapati masa pemprosesan tidak menunjukkan jurang yang besar. Namun untuk data yang melebihi 1000 baris, jelas menunjukkan bahawa algoritma jujukan dapat mempercepatkan masa pengekstrakan di antara 20% hingga 60%.



Rajah 5.2 : Perbandingan masa pengekstrakan dengan algoritma pengecam jujukan dan tanpa algoritma pengecam jujukan

Dengan merujuk kepada Rajah 6.2 di atas, peningkatan masa pemprosesan untuk algoritma EkstrakPro adalah secara eksponen. Ini kerana proses pengecaman

kata kunci perlu dilakukan untuk setiap data dalam fail input. Tetapi sekiranya menggunakan algoritma pengekstrakan maklumat bersama algoritma pengecaman jujukan, kadar penambahan masa dapat dipendekkan kerana data-data tidak perlu dibandingkan dengan kata kunci setelah mengetahui corak jujukan fail input.

Pengujian ini dapat membuktikan penggunaan algoritma pengecaman jujukan mampu mempercepatkan masa pengekstrakan terutamanya bagi data jujukan yang bersaiz besar.

5.5 Kesimpulan

Daripada kedua-dua pengujian yang telah dijalankan, dapat disimpulkan bahawa sistem EkstrakPro dapat meningkatkan ketahanan dan kepantasan proses pengekstrakan maklumat hidrologi JPS. Ketahanan sistem EkstrakPro berjaya membuktikan bahawa dengan penghasilan algoritma EkstrakPro yang berasaskan ontologi dapat mengekstrak format data hidrologi yang berlainan (sama ada format baru ataupun lama) dengan jayanya.

Apabila algoritma EkstrakPro digunakan bersama algoritma pengecaman jujukan, masa pengekstrakan maklumat dapat dikurangkan sebanyak 20% hingga 60% bagi data yang bersaiz besar. Manakala bagi perbandingan dengan menggunakan ukuran notasi-O pula adalah seperti berikut :

1. Algoritma EkstrakPro O(nm)
2. Algoritma EkstrakPro dengan algoritma pengecaman jujukan bagi
    - Kes Ideal O(n)
    - Kes Sederhana O(n) + h
    - Kes Tidak Ideal O(nm) + O(b)

Pengukuran notasi-O bagi kedua-dua algoritma ini jelas menunjukkan bahawa algoritma EkstrakPro menggunakan algoritma pengecaman jujukan mempunyai kiraan masa yang lebih rendah apabila keadaan kes adalah ideal dan sederhana.

BAB 6

KESIMPULAN

6.1 Pendahuluan

Penyelidikan ini telah membangunkan satu IE berasaskan ontologi bagi domain hidrologi yang telah mengambil kira masalah di JPS sebagai kajian kes. Kelemahan yang terdapat di dalam MHIS *Dataload* telah berjaya diselesaikan dengan menggunakan IE berasaskan ontologi yang dikenali sebagai EkstrakPro. Reka bentuk prototaip EkstrakPro yang diambil daripada kumpulan BYU telah disesuaikan dengan domain hidrologi. Penambahbaikan yang dilakukan adalah penyediaan ontologi pengekstrakan yang lebih sistematik dengan penggunaan UO. Pengenalan algoritma pengecaman jujukan telah memberi impak yang besar ke atas data hidrologi berjujukan, terutamanya bagi saiz data yang besar. Bab ini seterusnya akan membincangkan kelebihan dan kelemahan ke atas penambahbaikan sistem EkstrakPro yang telah dilakukan. Penambahbaikan untuk penyelidikan akan datang turut disertakan.

6.2 Rumusan Keseluruhan Penyelidikan

IE berasaskan ontologi telah dibuktikan dapat mengekstrak data hidrologi dengan baik. Jika sebelum ini IE berasaskan ontologi sering digunakan untuk mengekstrak halaman web pelbagai domain, penyelidikan kali ini telah membuka ruang bagi penggunaan IE berasaskan ontologi bagi domain hidrologi khusus untuk data hidrologi berjujukan.

Ontologi pengekstrakan dibina secara manual dan dengan adanya UO, pembinaan ontologi pengekstrakan akan lebih teratur. UO dapat membantu sekiranya terdapat maklumat baru yang ingin di ekstrak. Rangka UO boleh ditambah bila-bil masa ketika diperlukan tanpa mengkompil pengaturcaraan semula.

Ketepatan pengekstrakan juga berjaya dicapai kesan daripada algoritma EkstrakPro yang tegar. Set kata kunci yang lengkap dan terperinci juga menyumbang kepada ketepatan pengekstrakan data. Set kata kunci ini hanya dapat dihasilkan daripada ontologi pengekstrakan yang lengkap. Spesifikasi yang tidak lengkap akan menyebabkan pengekstrakan data yang tidak tepat. Penggunaan UO memudahkan penyenggaraan kata kunci dilakukan dari masa ke semasa.

Algoritma EkstrakPro turut mempunyai ketahanan terhadap perubahan struktur data. Ini kerana ianya menggunakan maklumat kata kunci dan tidak menggunakan maklumat lajur, jarak dan baris. Pengenalan data yang hendak di ekstrak adalah berdasarkan perkataan yang padan dengan kata kunci. Dengan kelebihan ini, data yang di ekstrak tidak perlu melalui proses *pre-cleaning* yang menyiah ralat struktur data.

Umumnya, perbandingan setiap perkataan di dalam data dengan kata kunci akan menyebabkan masa pemprosesan yang perlahan. Bagi mendapatkan masa pengekstrakan yang pantas, penulis telah menghasilkan algoritma pengecaman jujukan. Algoritma ini akan mengecam corak rekod di dalam data. Jika data mempunyai rekod yang berjujukan, langkah perbandingan perkataan dan kata kunci tidak dilaksanakan di dalam setiap baris rekod. Perkataan yang sedia dikenal pasti sebagai kata kunci daripada baris-baris sebelum akan di ekstrak tanpa memerlukan perbandingan dengan kata kunci. Ini terbukti dengan pengujian yang dijalankan. Pengekstrakan data dengan pengecam rekod jujukan menunjukkan masa pemprosesan yang lebih baik berbanding pengekstrakan tanpa pengecam rekod jujukan. Secara keseluruhannya, IE berasaskan ontologi bagi domain hidrologi telah berjaya dihasilkan

6.3 Kebaikan dan kelemahan Kajian

Algoritma EkstrakPro yang dihasilkan mempunyai kelebihan berbanding *MHIS Dataload*. Algoritma adalah lebih tahan atau tegar terhadap perubahan struktur data. Ketahanan dapat dicapai dengan menggunakan pendekatan ontologi. IE berasaskan ontologi menghasilkan set peraturan pengekstrakan daripada ontologi pengekstrakan. Objek dan hubungan di dalam spesifikasi digunakan sebagai set peraturan pengekstrakan. Set kata kunci yang dihasilkan daripada ontologi pengekstrakan adalah faktor utama ketahanan algoritma.

Set kata kunci ini dihasilkan secara automatik dapat memansuhkan beban pengguna mengemas kini peraturan pengekstrakan. Di samping mengendalikan setiap jenis data hidrologi, algoritma juga dapat mengendalikan setiap jenis format yang digunakan di JPS. Ini bermakna hanya satu algoritma digunakan bagi kerja pengekstrakan data hidrologi JPS. Algoritma lebih praktikal digunakan berbanding *MHIS Dataload* yang terdiri daripada 5 algoritma.

Oleh peraturan pengekstrakan berdasarkan set kata kunci dan bukan berdasarkan maklumat struktur seperti baris dan lajur data. Masa pemprosesan menjadi perlahan disebabkan algoritma melakukan perbandingan setiap perkataan di dalam baris data. Berbeza dengan *MHIS Dataload* yang melakukan kerja berdasarkan lajur yang dibaca. Melihat kepada langkah pengekstrakan, masa pemprosesan *MHIS Dataload* ternyata lebih pantas daripada algoritma EkstrakPro.

Namun penyelidikan ini telah menghasilkan algoritma tambahan yang dinamakan algoritma pengecaman jujukan bagi mengatasi masalah di atas. Hasil pengujian di dalam Bahagian 5.4 telah menunjukkan penggunaan bersama EkstrakPro dan algoritma pengecaman jujukan berjaya mengurangkan masa pemprosesan algoritma EkstrakPro.

Algoritma EkstrakPro memerlukan ontologi pengekstrakan berlainan sekiranya digunakan ke atas bidang yang berbeza. Ini adalah satu kelemahan di

dalam IE berasaskan ontologi. Sebagai contoh, algoritma yang menggunakan ontologi pengekstrakan bidang data hidrologi tidak dapat mengekstrak data daripada bidang jualan kereta kerana kata kunci yang digunakan berbeza.

6.4 Penambahbaikan

Penyelidikan telah mencapai matlamat di dalam menyelesaikan beban penyenggaraan pengekstrakan data MHIS dengan menghasilkan IE berasaskan ontologi. Walaupun EkstrakPro menunjukkan hasil yang memuaskan, namun masih terdapat ruang pembaikan yang boleh dilaksanakan. Pendekatan berasaskan ontologi memerlukan ontologi pengekstrakan yang dihasilkan secara manual. Kerja-kerja manual memerlukan seseorang yang mahir di dalam bidang tumpuan dan risiko ke atas kerja manual adalah kesilapan individu. Oleh kerana ketepatan pengekstrakan berkait langsung dengan terperincinya ontologi pengekstrakan, maka kerja ini wajar dilakukan secara automatik.

6.5 Penutup

Secara keseluruhannya laporan ini telah mengkaji kaedah pengekstrakan data daripada fail teks. Berdasarkan keputusan pengujian ke atas algoritma, pendekatan berasaskan ontologi yang diperkenalkan oleh kumpulan BYU, sesuai digunakan di dalam menyelesaikan masalah beban penyenggaraan pengekstrakan data MHIS dengan penambahan UO dan algoritma pengecaman jujukan. Secara keseluruhannya, penyelidikan telah berjaya mencapai matlamat dan juga objektif kajian.

BIBLIOGRAFI

Abiteboul, S. (1997). *Querying semi-structured data.* In Database Theory, 6th International Conference. January 8-10. Greece. 1-18.

Adelberg, B.(1998). *NoDoSE: A Tool for Semi-Automatically Extracting Structured and Semi-Structured Data from Text Documents*. SIGMOD Record 21(2): 283-294.

Arocena, G.O. and Mendelzon, A, O. (1998). *WebOQL: Restructuring Documents, Databases and Webs*. In Proceedings of the 14th IEEE International Conference on Data Engineering. Florida. 24-33.

Azmi Jafri (2002). *Arahan Kerja: Pemungutan data Hidrologi*. Technical Report. UPMBH-PK(AK)-01

Baumgartner, R., Sergio, F., Georg G. (2001). *Visual Web information extraction with Lixto.* In Proceedings of the 26th International Conference on Very Large Database Systems.119-128.

Brian, S., Rajeev M., Lawrence P., Terry W. (1998). *What can you do with a Web in your pocket?* Data Engineering Bulletin 21(2): 27-47.

Buneman, P. (1997). *Semi structure Data.* In Proceedings of the 16th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Arizona. 117-121.

Califf, M. E. and Mooney, R. J.(1999). *Relational learning of Pattern-Match Rules for*
Information Extraction. In Proceedings of 16th National Conference on Artificial Intelligence and 11th Conference on Innovative Applications of Artificial Intelligence.
. 328-334.

Cocchiarella, N.B. (1991). *Formal Ontology*. Handbook of Metaphysics and Ontology. Munich.

Crescenzi, V. and Mecca, G. (1998) Grammars Have Exceptions. *Information System* 23(8). 539-565.

Crescenzi, V., Giansalvatore, M., Paolo, M. (2001). *RoadRunner: Towards Automatic Data Extraction from Large Web Sites*. In Proceedings of the 26th International Conference. Italy. 109-118.

Embley D. W., Barry D. K, Scott N. W. (1992). *Object oriented Systems Analysis: A Model-Driven Approach*. Prentice Hall, Englewood, New Jersey.

Embley D. W., Douglas M. C., Stephen W. L., Randy D. S.(1998). *Ontology-based extraction and structuring of information from data-rich unstructured documents*. In Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM'98), pages 52–59.

Florescu, D., Alon Levy, Alberto Mendelzon (1998). *Database techniques for the World-Wide Web: A Survey*. SIGMOD Record: 27(3). 59-74.

Freitag, D. (2000). *Machine learning for information in Informal Domains*. Machine Learning 2/3. 162-2002.

Golgher, P. B., Altigran S. da Silva., Alberto H. F. Laender, Berthier Ribeiro-Neto (2001). *Bootstrapping for Example-Based Data Extraction*. In Proceedingof the 10th ACM International Conference on Information and Knowledge Management. Georgia.

Guarino, N. (1998). Some Ontological Principles for Designing upper level lexical resources. Proceeding of the first Iternational conference on lexical resources and evaluation. Granada.

Hammer, J., H. Garcia, M., S. N., R. Yerneni, M. M. Breunig, and V. V. (1997). *Template-Based Wrappers in the TSMMIS System*. SIGMOD Record: 26(2). 532-535.

Hammer, J. (1997). *The TSIMMIS Experience.* In Proceedings of the First East-European Symposium on Advances in Databases and Information Systems. Rusia. 1-8.

Hsu, C. n. and Dung, M. T. (1998). *Generating Finite-State Transducers for Semi-Structured Data Extraction from the web*. Information Systems 23(8). 521-538.

Huck, G., Peter F., Karl A., Erich N. (1998). *jedi: Extracting and Synthesizing Information from the web*. In Proceeding of the 3rd IFCIS International Conference on cooperative Information Systems. New York. 32-43.

Hwang, C.H. (1999). Incompletely and Imprecisely Speaking: Using Dynamic Ontologies for representing and retrieving information. Proceeding of 6th international workshop on knowledge representation meets databases, KRDB'99.Sweden.

Jabatan Pengairan dan Saliran Malaysia. (2001a). *Malaysian Hydrological Information System, Final Report. Vol 1*. Technical Report

Jabatan Pengairan dan Saliran Malaysia. (2001b). *Malaysian Hydrological Information System, Final Report. Vol 2.* Technical Report

Jabatan Pengairan dan Saliran Malaysia. (2001c). *Malaysian Hydrological Information System, Final Report. Vol 3.* Technical Report

Jabatan Pengairan dan Saliran Malaysia. (2001d). *Malaysian Hydrological Information System, Final Report. Vol 4*. Technical Report

Kushmerick, N. (2000). Wrapper Induction: Efficiency and Expressiveness. *Artificial Intelligence Journal* 118(1/2): 15-68.

Laender, A.H.F, Berthier, A, Ribeiro-Neto, da Silva, Altigran Soares (2001). DEByE – data Extraction by Example. Data and Knowledge Engineering.

Laender, A. H. F (2000). *Representing Wed Data as Complex Objects.* In Electronic Commerce and Web technologies. Berlin. 216-228.

Lopez, F. M.(1999). *Overview of Methodologies for Building Ontologies.* Proceeding of IJCAI-99 Workshop on Ontologies and Problem Solving Method. Stockholm.

Mecca, G., Atzeni, P., Masci, A., Merialdo, P., Sindoni, G (1998). The ARANEUS Web-base management System. *SIGMOD*
Record: 27(2). 544-546.

Muslea, I. (1999). *Extraction Pattern for Information Extraction tasks: A survey*. In Proceeding of the AAAI-99 Workshop on Machine Learning for Information Extraction. Florida. 1-6.

Muslea, I., Steven Minton, Craig A. Knoblock (2001). *Hierarchical Wrapper Induction for Semi-structured Information Sources*. Autonomous Agents and Multi-Agent. 4(1/2).

Ribeiro-Neto, B. A. (1999). *Extracting Semi-Structured Data Through Examples*. In Proceeding of the 8th ACM International Conference on Information and Knowledge management. Missouri. 94-101.

Reich, J.R.(1999). Ontological Design Patterns for The Integration of Molecular Biological Information. Proceeding of German Conference on Bioinformatic GCB'99.October Hannover. 156-166.

Sahuguet, A. and Azavant F. (2001). *Building Intelligent Web Application using Lightweight wrappers.* Data and Knowledge Engineering 36(3): 283-316.

Soderlan, S. (1999). *Learning Information Extraction Rules for Semi-Structure and Free Text*. Machine Learning 24 (1/3): 233-272.
Uschold, M., Gruninger, M.(1996). Ontologies: Principles, Method and Application. Knowledge Engineering Review 11(2). 93-155.

Xiaoying, G. and Mengjie, Z. (2004). A Knowledge Learning Approach toInformation Extraction From Multiple Text Based Web Site. *International Jurnal On Artificial Intelligence Tools*. Vol.13, No.3:721-738.

Youn, C.(1992) .Data Migration. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics,* Chicago, Illinois, USA, Volume 2, pages 1255-1258.

LAMPIRAN A

Contoh rangka unit objek bagi stesen ID

Rangka : Stesen_Id	
Nombor Unit Objek :1	
Bilangan item leksikel : 1	
No. Sub-rangka : 1 Style : number Pattern No: 1 Fungsi corak : Sebarang_digit Mak : 9999999 Min : 0000000	

Contoh unit objek stesen_id jenis "1632301**"**

Rangka : Stesen_Id	
Nombor Unit Objek :2	
Bilangan item leksikel : 3	
No. Sub-rangka : 1 Style : number Pattern No: 1 Fungsi corak : Sebarang_digit Mak : 9999999 Min : 0000000	No. Sub-rangka : 2 Style : frasa Intances : ["site"]
No. Sub-rangka : 3 Style : char Intances : [" "]	

Contoh unit objek stesen_id jenis "site 1732001**"**

Rangka : Stesen_Id	
Nombor Unit Objek :3	
Bilangan item leksikel : 2	
No. Sub-rangka : 1	No. Sub-rangka : 2

Style : number Pattern No: 1 Fungsi corak : Sebarang_digit Mak : 9999999 Min : 0000000	Style : frasa Intances : [“SA”, “-R”]

Contoh unit objek stesen_id jenis “SA-R4815001**”**

LAMPIRAN B

Contoh rangka unit objek bagi tarikh cerapan

Rangka : Tarikh_cerapan	
Nombor Unit Objek : 1	
Bilangan item leksikel : 5	
No. Sub-rangka : 1 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 2111 Min : 1960	No. Sub-rangka : 2 Style : char Intances : [“”]
No. Sub-rangka : 3 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 12 Min : 01	No. Sub-rangka : 4 Style : char Intances : [“”]
No. Sub-rangka : 5 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 31 Min : 01	

Contoh unit objek tarikh_cerapan jenis “20000104**”**

Rangka : Tarikh_cerapan	
Nombor Unit Objek : 2	
Bilangan item leksikel : 6	
No. Sub-rangka : 1 Style : frasa Intances : [“SRD-” , “ERD-”]	No. Sub-rangka : 2 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit

	Max : 31 Min : 01
No. Sub-rangka : 3 Style : char Intances : [“/”]	No. Sub-rangka : 4 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 12 Min : 01
No. Sub-rangka : 5 Style : char Intances : [“/”]	No. Sub-rangka : 6 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 2111 Min : 1960

Contoh unit objek tarikh_cerapan jenis “SRD-10/01/2002” dan “ERD-27/03/2002”

Rangka : Tarikh_cerapan	
Nombor Unit Objek : 1	
Bilangan item leksikel : 5	
No. Sub-rangka : 1 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 31 Min : 01	No. Sub-rangka : 2 Style : char Intances : [“-”]
No. Sub-rangka : 3 Style : frasa Intances : [“JAN” , “FEB” , “MAR” , “APR” , “MAY” , “JUN” , “JUL” , “AUG” , “SEP” , “OCT” , “NOV” , “DEC”]	No. Sub-rangka : 4 Style : char Intances : [“-”]

No. Sub-rangka : 5 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 2111 Min : 1960	

Contoh unit objek tarikh_cerapan jenis “31-DEC-2003” dan “1-JAN-2001”

LAMPIRAN C

Contoh rangka unit objek bagi masa cerapan

Rangka : Masa_cerapan	
Nombor Unit Objek : 1	
Bilangan item leksikel : 3	
No. Sub-rangka : 1 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 23 Min : 00	No. Sub-rangka : 2 Style : char Intances : [":"]
No. Sub-rangka : 3 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 59 Min : 00	

Contoh unit objek masa_cerapan jenis "09:25", "22:45","00:11"

Rangka : Masa_cerapan	
Nombor Unit Objek : 1	
Bilangan item leksikel : 4	
No. Sub-rangka : 1 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 12 Min : 00	No. Sub-rangka : 2 Style : char Intances : [":"]
No. Sub-rangka : 3 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit	No. Sub-rangka : 4 Style : frasa Intances : ["am" , "pm"]

Max : 59 Min : 00	

Contoh unit objek masa_cerapan jenis “8:00:00am”,”3:45:10pm”

Rangka : Masa_cerapan	
Nombor Unit Objek : 3	
Bilangan item leksikel : 3	
No. Sub-rangka : 1 Style : frasa Intances : [“SRT-” , “ERT-”]	No. Sub-rangka : 2 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 23 Min : 00
No. Sub-rangka : 3 Style : char Intances : [“:”]	No. Sub-rangka : 4 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 59 Min : 00
No. Sub-rangka : 5 Style : char Intances : [“:”]	No. Sub-rangka : 6 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 59 Min : 00

Contoh unit objek masa_cerapan jenis “SRT-14:30:15”,”ERT-16:21:52”

Rangka : Masa_cerapan	
Nombor Unit Objek : 4	
Bilangan item leksikel : 5	
No. Sub-rangka : 1 Style : digit Pattern No: 1	No. Sub-rangka : 2 Style : char Intances : [“”]

Fungsi corak : sebarang_digit Max : 23 Min : 00	
No. Sub-rangka : 3 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 59 Min : 00	No. Sub-rangka : 4 Style : char Intances : [“”]
No. Sub-rangka : 5 Style : digit Pattern No: 1 Fungsi corak : sebarang_digit Max : 59 Min : 00	

Contoh unit objek masa_cerapan jenis “122504”, “151722”

LAMPIRAN D

Contoh keratan data hidrologi kategori pertama

```
2324032      1          INCREMENTAL
0            20030324   160004
5            20030324   203800
5            20030324   203824
5            20030324   203838
5            20030324   203903
5            20030324   203929
5            20030324   203950
5            20030324   204004
5            20030324   204016
```

Keratan data 1

```
Date Time          Event (Tipping 0.5 mm)
03/24/03 16:00:04.0        0
03/24/03 20:38:00.0        1
03/24/03 20:38:24.5        2
03/24/03 20:38:38.0        3
03/24/03 20:39:03.5        4
03/24/03 20:39:29.0        5
03/24/03 20:39:50.0        6
03/24/03 20:40:04.0        7
03/24/03 20:40:16.5        8
```

Keratan data 2

```
2324032      1          INCREMENTAL
0            20030225   092314
0            20030226   080000
5            20030226   133726
5            20030226   134208
5            20030226   184023
5            20030226   184048
5            20030226   184208
5            20030226   184300
5            20030226   184329
```

Keratan data 3

```
"1313006","1","INCREMENTAL"
"6","790126","163903"
"110","790126","164828"
"250","790126","171829"
"126","790126","183825"
"89","790126","203410"
"11","790127","75748"
"0","790127","83000"
"3","790303","121702"
"26","790303","123132"
```

Keratan data 4

```
4815001        1 INCREMENTAL
    0    1020110    143015
    0    1020110    144045
    5    1020110    144046
    0    1020110    164227
    5    1020110    164228
    0    1020110    164855
    5    1020110    164856
    0    1020110    202516
    5    1020110    202517
```

Keratan data 5

```
Datalog Report  11
SA-R4815001
SRT-14:30:15-SRD-10/01/2002
ERT-16:21:52-ERD-27/03/2002
ENCT-00298
TIP- 0.5 mm
N00001D1001T144045N00002D1001T164227N00003D1001T164855N00004D1001T202516
N00005D1101T111119N00006D1201T014122N00007D1201T205939N00008D1201T210714
N00009D1201T212138N00010D1201T213749N00011D1201T215452N00012D1201T220331
N00013D1201T222503N00014D1201T224806N00015D1201T230620N00016D1201T231359
N00017D1201T234731N00018D1301T002413N00019D1501T110821N00020D1901T224850
N00021D2001T181449N00022D2001T203618N00023D2001T205815N00024D2001T211749
N00025D2101T024302N00026D2101T124109N00027D2101T1
```

Keratan data 6

LAMPIRAN E

Contoh keratan data hidrologi kategori kedua

```
23240321        INCREMENTAL
20030225        0       92314
20030226        0       80000
20030226        5       133726
20030226        5       134208
20030226        5       184023
20030226        5       184048
20030226        5       184208
20030226        5       184300
20030226        5       184329
```

Keratan data 1

```
23240321        INCREMENTAL
92314   20030225        0
80000   20030226        0
133726  20030226        5
134208  20030226        5
184023  20030226        5
184048  20030226        5
184208  20030226        5
184300  20030226        5
184329  20030226        5
```

Keratan data 2

```
131006,INCREMENTAL,
790126,163903,6
790126,164828,110
790126,171829,250
790126,183825,126
790126,203410,89
790127,75748,11
790127,83000,0
790303,121702,3
790303,123132,26
```

Keratan data 3

```
131006,INCREMENTAL,
790126,6,163903
790126,110,164828
790126,250,171829
790126,126,183825
790126,89,203410
790127,11,75748
790127,0,83000
790303,3,121702
790303,26,123132
```

Keratan data 4

```
131006,INCREMENTAL
Time 1600,
790126,66.9
790126,110
790126,300.6
790126,126
790126,89
790127,11.789
790127,0
790303,66.6
```

Keratan data 5

```
Date Time Event (Tipping)
38:00.0  1        3/24/2003
38:24.5  2        3/24/2003
38:38.0  3        3/24/2003
39:03.5  4        3/24/2003
39:29.0  5        3/24/2003
39:50.0  6        3/24/2003
40:04.0  7        3/24/2003
40:16.5  8        3/24/2003
40:29.0  9        3/24/2003
```

Keratan data 6

```
Date : 24/2/2004 Event Time
0       40:43.0
1       40:52.0
2       41:08.0
3       41:19.0
4       41:30.0
5       41:45.0
6       41:57.0
7       42:06.0
8       42:16.0
```

Keratan data 7

```
Date Time Event (Tipping)
38:00.0  1        3/24/2003
38:24.5  2        3/24/2003
38:38.0  3        3/24/2003
39:03.5  4        3/24/2003
39:29.0  5        3/24/2003
39:50.0  6        3/24/2003
40:04.0  7        3/24/2003
40:16.5  8        3/24/2003
40:29.0  9        3/24/2003
```

Keratan data 8

```
~~~ NIWA Tideda ~~~ JPS Hydrology and Water Resources Division
                                          22-MAY-2003 12:30
~~~ LIST ~~~
Source is Selangor.mtd Site 3516424 Sg. Selangor at Empang Pecah, Selangor
 1 Item   INSTANT     From 20030204 113000 to 20030305 130000
     Stage     Date    Time
       mm.

*** GAP ***
113000  50080    20030204
160611  50083    20030204
180316  50085    20030204
135505  50078    20030205
170112  50078    20030205
```

Keratan data 9

```
113000  50080.5 20030204
160611  50083.5 20030204
180316  50085   20030204
135505  5007.5  20030205
170112  50078   20030205
22235   50075   20030206
205320  50075   20030206
50540   50072.5 20030207
165709  50064   20030207
30034   50062.5 20030208
```

Keratan data 10

```
Stesen 345217
Date Time       Event (Tipping 0.5 mm)50080     113000  20030204
50083   160611  20030204
50085   180316  20030204
50078   135505  20030205
50078   170112  20030205
50075   22235   20030206
50075   205320  20030206
50072   50540   20030207
50064   165709  20030207
```

Keratan data 11

```
50080   113000  20030204
50083   160611  20030204
50085   180316  20030204
50078   135505  20030205
50078   170112  20030205
50075   22235   20030206
50075   205320  20030206
50072   50540   20030207
50064   165709  20030207
50072   50540   20030207
```

Keratan data 12

```
Stesen 345217
Date    Time Event (Tipping 0.5 mm)
50083   20030204
50085   20030204
50078   20030205
50078   20030205
50075   20030206
50075   20030206
50072   20030207
50064   20030207
```

Keratan data 13

```
131006,INCREMENTAL,
790303,121702,3.3
790303,123132,26
790304,80231,8
790304,82500,0
790126,163903,6.5
790126,164828,110.4
790126,171829,250.8
790126,183825,126
790126,203410,89.6
```

Keratan data 14

```
92314   20030225        0
80000   20030226        0
133726  20030226        5
134208  20030226        5
184023  20030226        5
184048  20030226        5
184208  20030226        5
184300  20030226        5
184329  20030226        5
184352  20030226        5
```

Keratan data 15

```
131006,INCREMENTAL
Time 1600,
790126,66.9
790126,110
790126,300.6
790126,126
790126,89
790127,11.789
790127,0
790303,66.6
```

Keratan data 16

```
131006,INCREMENTAL

790126,66.9
790126,110
790126,300.6
790126,126
790126,89
790127,11.789
790127,0
790303,66.6
```

Keratan data 17

```
StesenId 826317
Date      Event (Tipping 0.5 mm)
3/24/2003         0
3/24/2003         1
3/24/2003         2
3/24/2003         3
3/24/2003         4
3/24/2003         5
3/24/2003         6
3/24/2003         7
```

Keratan data 18

```
Date       Event (Tipping 0.5 mm)

3/24/2003          0
3/24/2003          1
3/24/2003          2
3/24/2003          3
3/24/2003          4
3/24/2003          5
3/24/2003          6
3/24/2003          7
```

Keratan data 19

```
~~~ NIWA Tideda ~~~ JPS Hydrology and Water Resources Division
                                        22-MAY-2003 12:30
~~~ LIST ~~~
Source is Selangor.mtd Site 3516424 Sg. Selangor at Empang Pecah, Selangor
 1 Item   INSTANT     From 20030204 113000 to 20030305 130000
     Stage     Date    Time
       mm.

*** GAP ***
113000  50080
160611  50083
180316  50085
135505  50078
170112  50078
22235    50075
```

Keratan data 20

```
23240321          INCREMENTAL
92314    Januari, 23, 1999 0
80000    Januari, 23, 1999 0
133726   Januari, 23, 1999 5
134208   Januari, 23, 1999 5
184023   Januari, 23, 1999 5
184048   Januari, 23, 1999 5
184208   Januari, 23, 1999 5
184300   Januari, 23, 1999 5
184329   Januari, 23, 1999 5
```

Keratan data 21

```
~~~ NIWA Tideda ~~~ JPS Hydrology and Water Resources Division
                                        22-MAY-2003 12:30
~~~ LIST ~~~
Source is Selangor.mtd Site 3516424 Sg. Selangor at Empang Pecah, Selangor
 1 Item   INSTANT     From 20030204 113000 to 20030305 130000
     Stage     Date    Time
       mm.

*** GAP ***
50080    113000  20030204
50083    160611  20030204
50085    180316  20030204
50078    135505  20030205
50078    170112  20030205
50075    22235   20030206
50075    205320  20030206
```

Keratan data 22

```
Date Time          Event (Tipping 0.5 mm)
03/24/03 16:00:04.0        0
03/24/03 20:38:00.0        1
03/24/03 20:38:24.5        2
03/24/03 20:38:38.0        3
03/24/03 20:39:03.5        4
03/24/03 20:39:29.0        5
03/24/03 20:39:50.0        6

03/24/03 20:40:04.0        7
```

Keratan data 23

```
Stesen 167263
Date 03107
Time 120000

6
110
250
126
89
11
0
3
26
8
```

Keratan data 24

```
790126,66.9
790126,110
790126,300.6
790126,126
790126,89
790127,11.789
790127,0
790303,66.6
790303,26
790304,8
```

Keratan data 25

```
790126,89
790127,11
790127,45.8
790303,3
790303,54.7
790304,8
790304,56.99

131006,INCREMENTAL
Time 1600
```

Keratan data 26

```
425123
Date      Event (Tipping 0.5 mm)

3/24/2003         0
3/24/2003         1
3/24/2003         2
3/24/2003         3
3/24/2003         4

3/24/2003         5
3/24/2003         6
3/24/2003         7
```

Keratan data 27

```
50080   20030204 113000
50083   20030204 160611
50085   20030204 180316
50078   20030205 135505
50078   20030205 170112
50075   20030206 022235
50075   20030206 205320
50072   20030207 050540
50064   20030207 165709
50062   20030208 030034
```

Keratan data 28

```
Stesen 167263
Date 03107
Time 120000

6
110
250
126
89
11
0
3
26
```

Keratan data 29

```
03/24/03 20:49:08.0     42
03/24/03 43
03/24/03 20:50:00.0     44
03/24/03 20:50:22.5     45
03/24/03 20:50:47.5     46
03/24/03 20:51:19.0     47
03/24/03 20:51:52.0     48
03/24/03 20:52:31.0     49
03/24/03 20:53:33.5     50
03/24/03 51
03/24/03 52
03/24/03 20:58:00.5     53
03/24/03 54
03/24/03 21:00:09.5     55
03/24/03 21:01:55.5     56
```

Keratan data 30

LAMPIRAN F

Contoh keratan data hidrologi kategori ketiga

2324032	1	
	INCREMENTAL	
0	20030324	160004
5	20030324	203800
5	20030324	203824
5	20030324	203838
5	20030324	203903
5	20030324	203929
5	20030324	203950
5	20030324	204004
5	20030324	204016
5	20030324	204029
5	20030324	204043
5	20030324	204052
5	20030324	204108
5	20030324	204119
5	20030324	204130
5	20030324	204145
5	20030324	204157
5	20030324	204206
5	20030324	204216
5	20030324	204226
5	20030324	204237
5	20030324	204249
5	20030324	204301
5	20030324	204313
5	20030324	204325
5	20030324	204339
5	20030324	204352
5	20030324	204406
5	20030324	204415
5	20030324	204426
5	20030324	204435
5	20030324	204443
5	20030324	204456
5	20030324	204507
5	20030324	204518
5	20030324	204533
5	20030324	204556
5	20030324	204646
5	20030324	204711
5	20030324	204743
5	20030324	204811
5	20030324	204842
5	20030324	204908
5	20030324	204934
5	20030324	205000
5	20030324	205022
5	20030324	205047
5	20030324	205119
5	20030324	205152
5	20030324	205231
5	20030324	205333
5	20030324	205524
5	20030324	205651
5	20030324	205800
5	20030324	205857
5	20030324	210009
5	20030324	210155
5	20030324	210352
5	20030324	210650
5	20030324	210930
5	20030324	211116
5	20030324	211338
5	20030324	212428
5	20030324	213405

0	20030325	80000
5	20030325	164825
5	20030325	164914
5	20030325	164947
5	20030325	165027
5	20030325	165102
5	20030325	165128
5	20030325	165158
5	20030325	165232
5	20030325	165315
5	20030325	165355
5	20030325	165429
5	20030325	165520
5	20030325	165704
5	20030325	165817
5	20030325	170006
5	20030325	170150
5	20030325	170232
5	20030325	170306
5	20030325	170336
5	20030325	170410
5	20030325	170459
5	20030325	170529
5	20030325	170606
5	20030325	170642
5	20030325	170704
5	20030325	170716
5	20030325	170731
5	20030325	170745
5	20030325	170759
5	20030325	170812
5	20030325	170830
5	20030325	170848
5	20030325	170901
5	20030325	170913
5	20030325	170925
5	20030325	170934
5	20030325	170946
5	20030325	170959
5	20030325	171012
5	20030325	171023
5	20030325	171035
5	20030325	171047
5	20030325	171059
5	20030325	171109
5	20030325	171120
5	20030325	171129
5	20030325	171141
5	20030325	171151
5	20030325	171200
5	20030325	171209
5	20030325	171217
5	20030325	171227
5	20030325	171236
5	20030325	171245
5	20030325	171256
5	20030325	171306
5	20030325	171316
5	20030325	171326
5	20030325	171336
5	20030325	171345
5	20030325	171352
5	20030325	171401
5	20030325	171411
5	20030325	171422
5	20030325	171433
5	20030325	171444
5	20030325	171452
5	20030325	171502
5	20030325	171511
5	20030325	171521
5	20030325	171531
5	20030325	171542
5	20030325	171553
5	20030325	171602

5	20030325	171612
5	20030325	171622
5	20030325	171631
5	20030325	171639
5	20030325	171647
5	20030325	171654
5	20030325	171702
5	20030325	171711
5	20030325	171719
5	20030325	171725
5	20030325	171732
5	20030325	171739
5	20030325	171746
5	20030325	171751
5	20030325	171758
5	20030325	171806
5	20030325	171814
5	20030325	171822
5	20030325	171829
5	20030325	171838
5	20030325	171846
5	20030325	171856
5	20030325	171903
5	20030325	171910
5	20030325	171917
5	20030325	171923
5	20030325	171929
5	20030325	171936
5	20030325	171943
5	20030325	171953
5	20030325	172005
5	20030325	172020
5	20030325	172038
5	20030325	172052
5	20030325	172106
5	20030325	172125
5	20030325	172144
5	20030325	172207
5	20030325	172227
5	20030325	172247
5	20030325	172312
5	20030325	172338
5	20030325	172417
5	20030325	172500
5	20030325	172541
5	20030325	172614
5	20030325	172649
5	20030325	172721
5	20030325	172753
5	20030325	172822
5	20030325	172854
5	20030325	172934
5	20030325	173013
5	20030325	173058
5	20030325	173148
5	20030325	173235
5	20030325	173325
5	20030325	173431
5	20030325	173531
5	20030325	173616

Contoh data hidrologi yang mempunyai 200 baris